

# Efficient and Effective Manual Corpus Annotation

## To Create Resources for Evaluation and Machine Learning

DISSERTATION

zur Erlangung des akademischen Grades

**Doktor der Technischen Wissenschaften**

eingereicht von

**Dipl.-Ing. Markus Zlabinger, BSc**

Matrikelnummer 00828324

an der Fakultät für Informatik  
der Technischen Universität Wien  
Betreuung: Prof. Dr. Allan Hanbury

Diese Dissertation haben begutachtet:

---

Gianluca Demartini

---

Massimo Poesio

Wien, 12. Mai 2021

---

Markus Zlabinger



Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.  
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.



# Efficient and Effective Manual Corpus Annotation

Optional Subtitle of the Thesis

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

**Doktor der Technischen Wissenschaften**

by

**Dipl.-Ing. Markus Zlabinger, BSc**

Registration Number 00828324

to the Faculty of Informatics

at the TU Wien

Advisor: Prof. Dr. Allan Hanbury

The dissertation has been reviewed by:

---

Gianluca Demartini

---

Massimo Poesio

Vienna, 12<sup>th</sup> May, 2021

---

Markus Zlabinger



Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.  
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

# Erklärung zur Verfassung der Arbeit

Dipl.-Ing. Markus Zlabinger, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 12. Mai 2021

---

Markus Zlabinger



Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.  
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

# Acknowledgements

Pursuing a Ph.D. degree is the largest project that I have ever worked on. It is a life-changing journey with ups and downs – during which many critical decisions need to be made. Completing this journey was made only possible by the tremendous support received from colleagues, friends, and family members.

An important companion during my Ph.D. journey was my supervisor Allan Hanbury. Allan always had an open ear and gave excellent advice whenever critical decisions had to be made. I am thankful for all the feedback he provided throughout the years, allowing me to learn, improve, and grow from his experience.

I also express my deepest appreciation to all the colleagues that helped and supported me. I thank Marta Sabou, Mete Sertkan, Navid Rekabsaz, and Sebastian Hofstätter for the collaboration on various research projects. I also like to thank all the members of our research group for the inspiring discussions we had during lunch breaks or *jour fixe* meetings: Aldo Lipani, Abdel Aziz, Linda Andersson, Mihai Lupu, Serwah Sabetghadam, Florina Piroi, Tobias Fink, and Alexandros Bampoulidis.

I thank my family for the love and support provided throughout my entire life: my parents Roswitha and Gerhard Zlabinger; my brother Stefan; my grandparents Vera Flicker, Ernst Rohrer, Johann Zlabinger, and Theresia Zlabinger.

Last but certainly not least, I thank my girlfriend Walpurga Friedl for her love, support, and understanding, especially during the stressful times right before paper deadlines.



Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.  
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.



# Kurzfassung

Manuell annotierte Textkorpora sind ein unverzichtbarer Bestandteil der Forschungsgebiete Information Retrieval (IR) und Natural Language Processing (NLP). Wir benötigen solche Korpora für die systematische Evaluierung und das überwachte maschinelle Lernen. Obwohl wir in vielen Situationen auf annotierte Korpora angewiesen sind, ist die Erstellung eines neuen Korpus eine aufwendige Prozedur, bei der Menschen als Annotierer die Texte eines Korpus durcharbeiten, um passende Labels zuzuweisen. Erschwerend kommt hinzu, dass Annotierer bei der Zuweisung von Labels fehleranfällig sind, insbesondere bei domänenspezifischen Annotationsaufgaben, da diese in der Regel nur von erfahrenen Experten korrekt durchführbar sind.

In dieser Dissertation werden neuartige Ansätze, die Annotierer dabei unterstützen Labels schneller und genauer zuzuweisen, vorgestellt. Die Ansätze werden auf Textannotationsaufgaben aus den Forschungsbereichen IR und NLP angewandt, wie z. B. Question-Answering und Named-Entity Recognition. Darüber hinaus liegt ein Fokus dieser Arbeit auf Annotationsaufgaben aus dem biomedizinischen Bereich, da diese aufgrund der komplexen Fachsprache dieser Domäne schwierig zu annotieren sind.

Die erste entwickelte Methodik ist der GROUP-WISE-Ansatz, der Annotierer dabei unterstützt Labels schnell zuzuweisen. Die Idee dieses Ansatzes ist, Texte vor der Annotation basierend auf ihrer semantischen Ähnlichkeit vorzugruppieren. Die Annotation einer Gruppe semantisch ähnlicher Texte, wie z.B. Fragen, Sätze oder Phrasen, kann zeitsparend sein, besonders wenn jeder Text einen ähnlichen Label erfordert. Wir evaluieren den GROUP-WISE-Ansatz für die Aufgaben Question-Answering und Named-Entity Recognition. Unsere Resultate zeigen, dass die Vorgruppierung die Zeiteffizienz des Annotationsverfahrens erheblich verbessert, ohne die Korrektheit der zugewiesenen Labels zu beeinträchtigen.

Die zweite entwickelte Methodik unterstützt Annotierer, die keine Experten sind, bei der Durchführung von domänenspezifischen Annotationsaufgaben. Annotierer werden in der Regel mit Anweisungen und Beispielen auf solche Aufgaben vorbereitet. Die Bereitstellung von Beispielen ist wichtig, allerdings sind diese oft nur global für eine Aufgabe definiert und sind eventuell bei der Annotation spezifischer Texte nicht nützlich. Um dieses Problem zu lösen, schlagen wir den Ansatz Dynamic EXamples for Annotation (DEXA) vor. Dabei werden Annotierer durch *dynamische Beispiele*, die dem aktuell zu annotierenden Text semantisch ähnlich sind, unterstützt. Wir evaluieren den DEXA-Ansatz

für die Annotation von Named-Entities in Sätzen von biomedizinischen Publikationen. Die dynamischen Beispiele werden automatisch aus einer kleinen Menge von Sätzen bezogen, welche zuvor von Experten annotiert wurden. Wir rekrutieren Annotierer von Amazons Crowdsourcing-Plattform Mechanical Turk und messen die Übereinstimmung der Crowdworker im Vergleich zu Experten. Unsere Ergebnisse zeigen, dass Crowdworker, die durch den DEXA-Ansatz unterstützt werden, signifikant höhere Übereinstimmungen mit den Experten erreichen als Crowdworker ohne diese Unterstützung.

Der DEXA und der GROUP-WISE Annotationsansatz verwenden eine unüberwachte Semantic Short-Text Similarity Methode (SSTS), um die Ähnlichkeit zwischen Fragen und Sätzen zu berechnen. Um eine effektive Methode für beide Ansätze zu identifizieren, evaluieren wir zehn unüberwachte SSTS-Methoden auf vier Benchmark-Datensätzen. Die Resultate dieser Evaluierung zeigen die hohe Effektivität von Methoden, die auf Wordembeddings und kontextualisierten Textembeddings basieren. Wir verwenden diese Methoden für den DEXA und den GROUP-WISE Annotationsansatz.

Diese Dissertation liefert neue Forschungserkenntnisse um in den Bereichen IR und NLP den Prozess der manuellen Korpusannotation zu verbessern. Die entwickelten Annotationsansätze ermöglichen die effektive und zeiteffiziente Erstellung neuer Ressourcen für systematische Evaluierung und überwachtes maschinelles Lernen. Darüber hinaus liefert diese Arbeit neue Erkenntnisse auf dem Gebiet der Semantic Short-Text Similarity, indem sie umfangreiche Evaluierungsergebnisse für verschiedene Methoden und Datensätze beschreibt. Diese Evaluierungsergebnisse sind für andere Forscher, welche eine effektive Methode für ihre Anwendungsfälle benötigen, nützlich.

# Abstract

Manually annotated text corpora are an indispensable part of Information Retrieval (IR) and Natural Language Processing (NLP). We depend on annotated corpora for evaluation and supervised machine learning. While we depend on annotated corpora in many situations, creating a new one is a time-consuming procedure. It involves annotators going through the texts of a corpus to assign labels. To make things worse, annotators might be inaccurate in assigning labels, especially for domain-specific annotation tasks, as these usually require expert annotators to be conducted accurately.

This thesis introduces novel methodologies to support annotators in assigning labels more quickly and accurately. The methodologies are applied to text annotation tasks related to the IR and NLP research area, such as question-answering and named-entity recognition. Furthermore, we consider annotation tasks specific to the biomedical domain, as these are difficult to annotate accurately due to the complex jargon of this domain.

The first introduced methodology is the GROUP-WISE annotation approach to support annotators in assigning labels quickly. We propose to pre-group texts based on their semantic similarity before being annotated. Annotating a group of semantically similar texts, such as questions, sentences, or phrases, can be time-saving, especially when each text requires similar labeling. We evaluate the GROUP-WISE approach for question-answering and named-entity recognition. Our results show that pre-grouping substantially improves the annotation procedure's time efficiency without harming accuracy.

The second proposed methodology is about supporting non-expert annotators in conducting domain-specific annotation tasks. Annotators are commonly prepared for such tasks with instructions and examples. Providing examples is essential; however, they are usually defined globally over an entire task and might not be useful in labeling individual texts. We propose the Dynamic EXamples for Annotation (DEXA) approach, in which annotators are supported by *dynamic examples* semantically similar to the currently labeled text. We evaluate the DEXA approach for annotating named-entities in sentences of biomedical publications. We retrieve dynamic examples automatically from a small set of sentences previously labeled by experts. We recruit annotators from Amazon's Mechanical Turk crowdsourcing platform and measure their inter-annotator agreement to experts. Our results show that crowdworkers supported by the DEXA approach reach significantly higher agreements to the experts than crowdworkers without such support.

The DEXA and the GROUP-WISE annotation approach incorporate an unsupervised semantic short-text similarity method (SSTS) to compute the similarity between questions and sentences. To identify an effective method for our use case, we evaluate ten unsupervised SSTS methods on four benchmark datasets. Our results show the high effectiveness of methods based on word embeddings and contextualized text embeddings – which we use for the DEXA and the GROUP-WISE annotation approach.

This work contributes to improving the manual corpus annotation procedure for tasks related to the IR and NLP research area. The proposed methodologies allow researchers to create new resources for evaluation and supervised machine learning effectively and time-efficiently. Furthermore, the thesis contributes to the area of semantic short-text similarity by reporting extensive evaluation results for various methods and datasets. These results benefit other researchers seeking an effective method for their use cases.

# Contents

|   |             |
|---|-------------|
| <b>Kurzfassung</b>  | <b>ix</b>   |
| <b>Abstract</b>   | <b>xi</b>   |
| <b>Contents</b>   | <b>xiii</b> |
| <b>1 Introduction</b>   | <b>1</b>    |
| 1.1 Motivation and Challenges of Text-Annotation . . . . .        | 2           |
| 1.2 Thesis Goal . . . . .   | 5           |
| 1.3 Research Questions . . . . .                                  | 5           |
| 1.4 Contributions . . . . .                                       | 6           |
| 1.5 Published Research . . . . .                                  | 8           |
| 1.6 Thesis Structure . . . . .                                    | 10          |
| <b>2 Background on Manual Corpus Annotation</b>                   | <b>11</b>   |
| 2.1 Project Definition . . . . .                                  | 12          |
| 2.2 Data Preparation . . . . .                                    | 16          |
| 2.3 Execution . . . . .   | 19          |
| 2.4 Evaluation . . . . .  | 22          |
| 2.5 Examples of Annotation Projects . . . . .                     | 27          |
| 2.6 Summary . . . . .   | 41          |
| <b>3 State-of-the-Art</b>   | <b>43</b>   |
| 3.1 Efficient Text Annotation . . . . .                           | 43          |
| 3.2 Effective Text Annotation . . . . .                           | 45          |
| 3.3 Unsupervised Text Similarity Methods . . . . .                | 46          |
| 3.4 Summary . . . . .   | 51          |
| <b>4 Evaluation of Unsupervised Short-Text Similarity Methods</b> | <b>53</b>   |
| 4.1 Unsupervised Short-Text Similarity Methods . . . . .          | 54          |
| 4.2 Preprocessing and Pre-trained Language Models . . . . .       | 57          |
| 4.3 Medical Sentence Similarity . . . . .                         | 58          |
| 4.4 Similar Question Retrieval . . . . .                          | 59          |
| 4.5 Summary . . . . .   | 63          |
|   | xiii        |

|          |  |            |
|----------|--|------------|
| <b>5</b> | <b>Efficient Group-Wise Data Annotation</b>  | <b>65</b>  |
| 5.1      | Group-Wise Annotation Approach . . . . .   | 66         |
| 5.2      | Question-Answer Annotation . . . . .   | 67         |
| 5.3      | Biomedical Named-Entity Annotation . . . . .   | 72         |
| 5.4      | Summary . . . . .  | 77         |
| <b>6</b> | <b>Supporting Non-Experts for Complex Annotation Tasks</b>                             | <b>79</b>  |
| 6.1      | Dynamic Examples for Annotation . . . . .  | 80         |
| 6.2      | PIO Annotation Task . . . . .  | 80         |
| 6.3      | Experiment Setup . . . . .   | 83         |
| 6.4      | Results and Discussion . . . . .   | 84         |
| 6.5      | Summary . . . . .  | 90         |
| <b>7</b> | <b>Conclusion</b>  | <b>91</b>  |
| 7.1      | Research Questions and Contributions . . . . .   | 91         |
| 7.2      | Future Research . . . . .  | 93         |
| <b>A</b> | <b>Annotation Guidelines</b>   | <b>97</b>  |
| A.1      | Query-Document Relevance Labeling . . . . .  | 98         |
| A.2      | Biomedical Named-Entity Annotation . . . . .   | 101        |
| A.3      | Clinical Study Polarity Analysis . . . . .   | 113        |
| A.4      | Disease-Symptom Relevance Assessment . . . . .   | 118        |
| A.5      | Labeling of Age, Gender, Symptom in Case Reports . . . . .                             | 119        |
| A.6      | Labeling of Participant, Intervention, and Outcome in Clinical Trial Reports . . . . . | 120        |
|          | <b>List of Figures</b>   | <b>123</b> |
|          | <b>List of Tables</b>  | <b>125</b> |
|          | <b>Bibliography</b>  | <b>127</b> |

# Introduction

*We're entering a new world in which data may be more important than software.* – Tim O'Reilly

Datasets are a crucial resource to advance computer science. In 2009, the ImageNet dataset was released for the task of visual object recognition [DDS<sup>+</sup>09]. The dataset contains 14 million images, each manually assigned with labels out of 20,000 categories. The dataset fostered the development of ground-breaking methodologies used nowadays in fields and domains beyond image classification. For example, Krizhevsky et al. proposed in 2012 the Convolutional Neural Network (CNN) AlexNet [KSH12]. AlexNet reached a classification accuracy of 85% on the ImageNet dataset, outperforming the state-of-the-art by more than 10 percentage points [KSH12, RDS<sup>+</sup>15]. Accuracy further improved over the years, where recent work reports more than 95% accuracy, approaching human performance in conducting the task [CLX<sup>+</sup>17]. The tremendous advance of visual object recognition is often contributed to the ImageNet dataset [KSH12].

The key component of the ImageNet dataset is arguably the labels manually assigned to the 14 million images. Assigning such meta-information to data, be it images, videos, or texts, is known as *annotation*, and the persons who assign the labels are known as the *annotators*<sup>1</sup>. Datasets augmented by annotations (referred to as *annotated dataset*<sup>2</sup>) are used to evaluate the performance of new methodologies. The performance is measured based on metrics that allow us to determine the state-of-the-art supported by empirical evidence. Apart from evaluation, we require annotated datasets for the training and testing of supervised machine learning algorithms. Machine learning algorithms aim to automate tasks by learning patterns from the annotated data. Annotated datasets

---

<sup>1</sup>Although annotations can be computed and assigned by an automatic system, this thesis is exclusively concerned with *manual* annotation.

<sup>2</sup>For conciseness, we refer to an *annotated dataset* also as dataset.

for machine learning and systematic evaluation are publicly available. However, when we require annotations specific to a particular domain (e.g., biomedical, legal), task (e.g., question-answering, sentiment analysis), or language, we might experience a lack of annotated data. In case no appropriate dataset is available, a new one can be created through manual annotation.

Manually annotating data is usually a tedious, expensive, and time-consuming procedure. Furthermore, depending on the difficulty of an annotation task, annotators might assign incorrect labels. Obtaining high-quality labels is even more difficult for domain-specific tasks that require annotators with certain expertise to be conducted sufficiently well, such as tasks related to the legal, patent, or medical domain [ALP<sup>+</sup>16]. The quality and the size of a dataset dictate its utility for evaluation and supervised learning. Therefore, persons in charge of creating a new dataset—referred to as the *annotation practitioners*—pursue two main objectives: The annotation process should be efficient (in terms of time and cost) and effective (in terms of annotators conducting the task accurately).

In this thesis, we study the efficiency and effectiveness of the manual annotation procedure. Specifically, we propose new methodologies to support annotators in assigning labels more quickly and accurately. We set the research scope to text-annotation tasks of the Natural Language Processing (NLP) and Information Retrieval (IR) research area. Note that in the context of NLP, a dataset is commonly referred to as a *corpus*, and we use the terms *dataset* and *corpus* interchangeable throughout this thesis.

The remainder of this chapter is structured as follows: Section 1.1 describes our motivation and discusses the challenges of an efficient and effective annotation. We describe the thesis' goal in Section 1.2, the addressed research questions in Section 1.3, and the contributions in Section 1.4. We list our published work in Section 1.5 and give a road-map to the thesis chapters in Section 1.6.

### 1.1 Motivation and Challenges of Text-Annotation

While annotated corpora are indispensable for systematic evaluation and supervised learning, the underlying procedure of manual annotation is a niche research topic, with only a few IR and NLP researchers aiming to study and improve it. Consequently, many research directions remain unexplored on topics such as improving task design, annotator training, and annotator recruitment. Papers introducing new corpora usually use the standard annotation procedure, consisting at its core of the following steps [PS12a]:

1. defining the goal for collecting annotations
2. assembling the corpus from raw text data
3. training the annotators using instructions and examples
4. annotating the corpus
5. obtaining and evaluating the annotations



We conducted various annotation projects in our research group following the aforementioned traditional annotation procedure. During these projects, we encountered two core challenges common to annotation tasks: The first challenge is about the annotation procedure's efficiency, which is usually time- and cost-inefficient. The second challenge is about the procedure's effectiveness since creating a high-quality annotated corpus is often difficult [Wal18a, DKC<sup>+</sup>18]. We addressed these challenges by adapting the traditional annotation procedure with the aim of improving its efficiency and effectiveness. Motivated by promising preliminary results, we decided to improve our methodologies further and perform thorough systematic evaluations. The methodologies and systematic evaluations represent the foundation of this thesis.

### 1.1.1 Time- and Cost-Efficiency

The trend in corpus annotation is shifting towards labeling large volumes. For example, the Microsoft MACHine Reading COMprehension dataset (MS MARCO) contains manual annotations for more than one million search queries [BCC<sup>+</sup>16]. Another well-known dataset is the Stanford Question Answering Dataset (SQUAD), comprising more than 100,000 annotated questions [RZLL16]. Creating large datasets is important to saturate the data needs of complex machine learning architectures, such as deep learning networks. Furthermore, larger datasets benefit evaluation since results are more reliable and statistical significance tests more expressive. In general, the more annotations collected, the better, as long as acquiring additional annotations does not come at the cost of reduced label quality [PS12e].

The size of a corpus can be defined based on the number of *text samples* comprised. The concrete type of a text sample depends on the task and can be, e.g., e-mails (e.g., for spam/no-spam annotation), questions (e.g., for question-answering), or words (e.g., for part-of-speech tagging). Text samples are distributed to human workers who process each cognitively to decide what label should be assigned. Cognitive processing can be time-consuming, depending on the experience of the annotator, the difficulty of the task, and the complexity of the currently labeled sample. The annotators expect in return for the conducted labor compensation, which is usually monetary. The expenses for paying annotators correlate with the corpus size, as labeling larger corpora requires more resources such as time, money, and annotators. Further costs accumulate when hiring expert annotators for domain-specific tasks.

### 1.1.2 Effective Acquisition of High-Quality Labels

Another challenge of the manual annotation procedure is to create a dataset of high-quality. The quality of a dataset depends on the correctness of the underlying labels assigned by the annotators. The ability of annotators to conduct a task sufficiently well depends on various criteria, including:

- the task design
- the difficulty of the task
- the training of annotators with instructions and examples
- the qualification of annotators with respect to their experience, profession, and background knowledge

A commonly used metric to determine the quality of an annotated dataset is the inter-annotator agreement. The inter-annotator agreement measures the agreement and disagreement between annotators for conducting a task and is computed based on text samples redundantly labeled by several annotators. A high inter-annotator agreement is desirable as it shows that humans agree on how the task should be performed, a critical prerequisite for automation. Overall, the inter-annotator agreement indicates the quality and reliability of an annotated corpus for evaluation and supervised machine learning [McH12].

Creating high-quality datasets is especially challenging for so-called *expert tasks*, which require annotators with several years of knowledge or a certain profession to be conducted correctly [XY12]. This thesis studies expert tasks specific to the biomedical domain. Acquiring annotations for biomedical tasks is difficult due to the terminology and jargon that prevails in the medical literature [Wal18a, HLD06]. For example, consider the sentence presented in Table 1.1 and the task of labeling the sentence’s polarity as either positive or negative<sup>3</sup>. To assess the presented sentence’s polarity correctly, the annotator needs to know that the abbreviation PONV stands for Postoperative Nausea and Vomiting, and the reduction of PONV, therefore, suggests a positive polarity.

A common approach to address the difficulty of domain-specific annotation tasks is to recruit *expert annotators*. These have a profound education, background, and experience to conduct a domain-specific annotation task accurately. A disadvantage in recruiting experts is that they are costly, and their recruitment is tedious due to the scarce availability. Alternatively, we can recruit *non-expert* annotators who are cost-efficient and highly available compared to experts [SBDS14]. However, the recruitment of non-experts usually reduces the quality and reliability of the assigned annotations, as they lack the qualification to conduct an expert task sufficiently well [NLP<sup>+</sup>18].

---

<sup>3</sup>We describe and perform the medical polarity annotation task in Chapter 2 of this thesis.

Table 1.1: Example of the medical polarity annotation task. The aim of this task is to assess whether the sentence has a positive or negative polarity. The presented sentence is from the abstract of [ORLS11].

---

Treatment of established PONV comprising ondansetron and droperidol, with or without dexamethasone, reduced PONV in both treatment groups.

---

## 1.2 Thesis Goal

*This thesis aims to improve the manual text annotation procedure with respect to effectiveness and time efficiency.* We study IR and NLP-related tasks to improve the data annotation procedure for these research areas. The three main research aims of this thesis can be summarized as follows:

- The first aim is to improve the annotation procedure’s time efficiency, allowing practitioners to obtain more labeled data with their available budget.
- The second aim is to improve the annotation procedure’s effectiveness by increasing the accuracy of non-expert annotators for conducting domain-specific tasks.
- The third aim is to foster the general understanding of manual text annotation by analyzing various tasks and their specific challenges.

## 1.3 Research Questions

The first research question addresses the time efficiency of the manual annotation procedure. The manual annotation procedure is carried out by annotators who label each text sample (e.g., a question, e-mail, or sentence) of a corpus. This procedure is inefficient, requiring the annotator to cognitively process each sample one by one to assign the most suitable label. A more time-efficient alternative might be the pre-grouping of samples based on their semantic similarity. Annotating a group of semantically similar samples can be time-saving, especially when each requires similar labeling. A task that could benefit from pre-grouping is question-answering since the same questions might occur frequently and require annotating the same answer label. Besides question-answering, we investigate whether pre-grouping also benefits the task of labeling named-entities in sentences of medical publications. These sentences often have a similar phrasing, and their pre-grouping might speed-up the labeling procedure. We define the first research question of this thesis; RQ1:

*How does pre-grouping of similar text samples impact the annotators’ time efficiency for question-answering and biomedical named-entity recognition?*

The second research question is about effectively collecting high-quality annotations for domain-specific tasks. Recruiting expert annotators for domain-specific tasks is expensive and cumbersome due to their limited availability. An alternative is recruiting annotators from crowdsourcing platforms such as Amazon’s Mechanical Turk. Although these *crowdworkers* are cost-efficient and highly available, they usually lack the expertise to conduct domain-specific tasks with high accuracy [Wal18a]. A common approach to compensate for the lack of expertise is to provide task instructions and examples that demonstrate how the task should be performed. Providing examples is essential; however,

they are usually defined globally over an entire task and might not be helpful in annotating individual samples. Examples that are semantically similar to the currently annotated sample—so-called *dynamic examples*—might provide better support to annotators and help them assigning accurate labels more often. We evaluate the effectiveness of dynamic examples for the task named-entity annotation in sentences of biomedical publications. We define the second research question of this thesis; RQ2:

*How does showing examples similar to the currently annotated text sample—so-called dynamic examples—affect the label accuracy of non-expert annotators for biomedical named-entity recognition?*

The defined research questions RQ1 and RQ2 compute the semantic similarity between text samples. For RQ1, we compute the semantic similarity between questions and sentences so that these can be pre-grouped. For RQ2, we compute the semantic similarity between sentences to retrieve dynamic examples. As sentences and questions are usually short texts, we require an effective semantic short-text similarity method (SSTS) to answer RQ1 and RQ2. Moreover, the SSTS method must be unsupervised since we aim to create annotated data independent of external corpora for supervised training.

The last addressed research question is about evaluating the effectiveness of unsupervised SSTS methods for question-answering and biomedical sentence retrieval. There are various methods available, ranging from classical methods such as computing the term frequency-inverse document frequency (TFIDF) [BR99] to more recently proposed methods based on word embeddings and contextualized text embeddings [PGJ18, RG19]. Although evaluation results are available occasionally for some methods and datasets, the literature lacks a systematic evaluation of unsupervised SSTS methods for question-answering and biomedical sentence retrieval. Hence, we define the third research question of this thesis; RQ3:

*How effective are (i) traditional, (ii) embedding-based, and (iii) contextualized unsupervised semantic short-text similarity methods for question-answering and biomedical sentence retrieval?*

## 1.4 Contributions

We now summarize our contributions related to the defined research questions. The first contribution is about improving the annotation procedure’s time efficiency by pre-grouping similar text samples (RQ1). We propose the GROUP-WISE annotation approach in which semantically similar samples are pre-grouped to speed-up cognitive processing. We compare the GROUP-WISE approach to the traditional approach of labeling each sample one by one, referred to as the SEQUENTIAL approach. We consider two tasks for our systematic evaluation: biomedical named-entity recognition and customer-support question-answering. For each task, we recruit annotators and assign them to use either

the GROUP-WISE or SEQUENTIAL approach. We compare the annotators' efficiency for metrics such as time, the number of clicks, and the number of interactions. As an additional metric, we analyze the inter-annotator agreement to determine a possible effect on the label quality when using either approach. Our results show that the GROUP-WISE approach can be used to assign annotations more time- and cost-efficiently compared to the SEQUENTIAL approach by preserving the same label quality.

The next contribution is about improving the effectiveness of non-expert annotators for conducting domain-specific annotation tasks (RQ2). We propose a new annotation approach in which we support non-expert annotators with dynamic examples that are similar to the currently labeled text sample. We refer to this annotation approach as Dynamic EXamples for Annotation (in short, DEXA). We evaluate the DEXA approach for labeling named-entities in sentences extracted from biomedical publications. For our evaluation, we recruit crowdworkers and divide them into two groups. The first group is provided task instructions and a few examples that demonstrate how the task should be performed. The second group is additionally supported by dynamic examples using the DEXA approach. The dynamic examples are automatically retrieved via a semantic text similarity method from a small set of samples previously labeled by experts. We measure the annotation quality by computing the inter-annotator agreement to a gold standard, consisting of annotations of medical experts. Our results show that crowdworkers supported by the DEXA approach reach significantly higher agreements to the experts than crowdworkers without such support.

The next contribution of this thesis is about the evaluation of unsupervised semantic short-text similarity methods (RQ3). We require an effective SSTS for the DEXA and the GROUP-WISE approach to compute the similarity between questions and sentences. We consider four benchmark datasets for our evaluation: Two biomedical short-text similarity datasets [SÖÖ17, WAF<sup>+</sup>18] and two question-to-question similarity datasets [NHM<sup>+</sup>17]. Based on these datasets, we evaluate the effectiveness of ten unsupervised SSTS methods, ranging from traditional word-based models such as TFIDF to more recent contextualized embedding methods such as Sentence-BERT [RG19]. Our results show the high effectiveness of methods based on sentence embeddings and word embeddings. Based on these results, we use the Sent2Vec method [PGJ18] to compute the similarity between biomedical sentences and the Smooth Inverse Frequency (SIF) method [ALM17] to compute the similarity between questions. The performed evaluation benefits other researchers seeking an effective SSTS method for their use cases.

As an additional contribution of this thesis, we create several new annotated corpora: We describe the fundamental framework for corpus annotation in Chapter 2, where we also demonstrate the framework's application for collecting annotations for five tasks related to IR and NLP. Chapter 5 extends the framework using the GROUP-WISE annotation approach to create new annotated corpora for question-answering and biomedical named-entity recognition. Finally, in Chapter 6, we use the DEXA approach to create an annotated corpus for named-entity recognition in biomedical publications.

An overview of the created datasets is given in Table 1.2, with details available in the referenced chapters. We make the datasets publicly accessible except for the customer-support question-answering dataset due to restrictions of the data provider. Links to access and download the datasets are available in the *Summary* section of each chapter. By publishing these datasets, we provide crucial resources for evaluation and supervised machine learning. Moreover, the datasets are relevant for researchers studying inter-annotator behavior, as all texts are labeled redundantly by multiple annotators.

## 1.5 Published Research

This thesis is heavily based on published work in IR and NLP-related journals and conferences. The following published papers are the foundation of this thesis:

- Medical Entity Corpus with PICO elements and Sentiment Analysis. **Conference on Language Resources and Evaluation (LREC) 2018**. *M Zlabinger, L Andersson, A Hanbury, M Andersson, V Quasnik, J Brassey* [ZAH<sup>+</sup>18]
- Extracting the Population, Intervention, Comparison and Sentiment from Randomized Controlled Trials. **Conference on Medical Informatics Europe (MIE) 2018**. *M Zlabinger, L Andersson, J Brassey, A Hanbury* [ZABH18]

Table 1.2: Overview of the annotated datasets created in the scope of this thesis. The labeled text sample are, depending on the task, pairs, questions, sentences, or abstracts of biomedical publications. Multiple annotators labeled each sample to allow the analysis of cross-annotator behavior, such as the inter-annotator agreement.

| Annotation Task   | Corpus Size     | ∅Annotations per Sample |
|---|-----------------|-------------------------|
| Query-document relevance assessment (Ch. 2)   | 24,199 pairs    | 3                       |
| Query-document relevant text span labeling (Ch. 2)  | 24,199 pairs    | 3                       |
| Biomedical named-entity annotation of <i>Participants</i> , <i>Interventions</i> , and <i>Comparisons</i> (Ch. 2) | 1,416 abstracts | 2                       |
| Clinical study polarity analysis (Ch. 2)  | 1,147 abstracts | 2                       |
| Disease-symptom relevance assessment (Ch. 2)  | 232 pairs       | 3                       |
| Customer-support question-answering (Ch. 5)   | 500 questions   | 2                       |
| Biomedical named-entity annotation of <i>Age</i> , <i>Gender</i> , and <i>Symptom</i> (Ch. 5)                     | 90 sentences    | 10                      |
| Biomedical named-entity annotation of <i>Participants</i> , <i>Interventions</i> , and <i>Outcomes</i> (Ch. 6)    | 423 sentences   | 6                       |



- Developing a fully automated evidence synthesis tool for identifying, assessing and collating the evidence. **British Medical Journal on Evidence-Based Medicine (BMJ) 2019.** *J Brassey, C Price, J Edwards, M Zlabinger, A Bampoulidis, A Hanbury [BPE<sup>+</sup>19]*
- DSR: A Collection for the Evaluation of Graded Disease-Symptom Relations. **European Conference on Information Retrieval (ECIR) 2020.** *M Zlabinger, S Hofstätter, N Rekabsaz, A Hanbury [ZHRH20]*
- Fine-Grained Relevance Annotations for Multi-Task Document Ranking and Question Answering. **Conference on Information & Knowledge Management (CIKM) 2020.** *S Hofstätter, M Zlabinger, M Sertkan, M Schröder, A Hanbury [HZS<sup>+</sup>20]*
- Improving the Annotation Efficiency and Effectiveness in the Text Domain. **European Conference on Information Retrieval (ECIR) 2019.** *M Zlabinger [Zla19b]*
- Efficient and Effective Text-Annotation Through Active Learning. **SIGIR Conference on Research and Development in Information Retrieval (SIGIR) 2019.** *M Zlabinger [Zla19a]*
- Efficient Answer-Annotation for Frequent Questions. **Conference of the Cross-Language Evaluation Forum (CLEF) 2019.** *M Zlabinger, N Rekabsaz, S Zlabinger, A Hanbury [ZRZH19]*
- DEXA: Supporting Non-Expert Annotators with Dynamic Examples from Experts. **SIGIR Conference on Research and Development in Information Retrieval (SIGIR) 2020.** *M Zlabinger, M Sabou, S Hofstätter, M Sertkan, A Hanbury [ZSH<sup>+</sup>20]*
- Effective Crowd-Annotation of Participants, Interventions, and Outcomes in the Text of Clinical Trial Reports. **Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020.** *M Zlabinger, S Hofstätter, M Sertkan, A Hanbury [ZSHH20]*

Other published papers not directly related to this thesis are the following:

- Mitigating the Position Bias of Transformer Models in Passage Re-Ranking. **European Conference on Information Retrieval (ECIR) 2021.** *S Hofstätter, A Lipani, S Althammer, M Zlabinger, A Hanbury [HLA<sup>+</sup>21]*
- Interpretable & Time-Budget-Constrained Contextualization for Re-Ranking. **European Conference on Artificial Intelligence (ECAI) 2020.** *S Hofstätter, M Zlabinger, A Hanbury [HZH20a]*

- Neural-IR-Explorer: A Content-Focused Tool to Explore Neural Re-ranking Results. **European Conference on Information Retrieval (ECIR) 2020**. *S Hofstätter, M Zlabinger, A Hanbury* [HZH20b]
- Verifying Extended Entity Relationship Diagrams with Open Tasks. **Conference on Human Computation and Crowdsourcing (HCOMP) 2020**. *M Sabou, K Käsžnar, M Zlabinger, S Biffl, D Winkler* [SKZ<sup>+</sup>20]
- Learning to Re-Rank with Contextualized Stopwords. **Conference on Information & Knowledge Management (CIKM) 2020**. *S Hofstätter, A Lipani, M Zlabinger, A Hanbury* [HLZH20]
- TU Wien@ TREC Deep Learning'19–Simple Contextualization for Re-ranking. **Text REtrieval Conference (TREC) 2019**. *S Hofstätter, M Zlabinger, A Hanbury* [HZH19]
- Finding Duplicate Images in Biology Papers. **Symposium on Applied Computing (SAC) 2017**. *M Zlabinger, A Hanbury* [ZH17]

### 1.6 Thesis Structure

The thesis is structured as follows: Chapter 2 describes the commonly used corpus annotation framework, and we evaluate the framework by acquiring annotations for five tasks related to the IR and NLP research domain. Chapter 3 reviews related work on (i) time-efficient annotation, (ii) effective annotation, and (iii) unsupervised short-text similarity. In Chapter 4, we conduct the comparative evaluation of the ten unsupervised semantic short-text similarity methods. Based on this evaluation, we select an effective method for the GROUP-WISE and the DEXA annotation approach, proposed in Chapters 5 and 6, respectively. We conclude the thesis in Chapter 7 by summarizing the main findings and discussing future research opportunities for manual corpus annotation.



# Background on Manual Corpus Annotation

Creating a new annotated corpus is a complex procedure [FE17]. The procedure involves activities such as acquiring the raw text data, recruiting suitable annotators, and evaluating the quality of the annotations. The annotation practitioners manage these activities and need to make various critical decisions throughout the annotation procedure. Furthermore, they need to use their available resources (e.g., time, budget) in the best way possible to create an annotated corpus of sufficient size and quality.

Best practices on manual corpus annotation are available to support practitioners with their decisions. Leech [Lee05] describes standards and best practices for various aspects of the annotation procedure, such as training the annotators, using specialized software to support annotators, and measuring the annotations' quality. McEnery et al. [MXT06a] provide a manual for corpus annotation. The manual focuses on best practices for traditional linguistic tasks, such as part-of-speech tagging, lemmatization, and coreference annotation. Although published more than a decade ago, the listed literature [Lee05, MXT06a] teaches the basics of corpus annotation, still essential today.

More recent work describes best practices for new directions of manual corpus annotation. Sabou et al. [SBDS14] describe best practices for scaling annotation projects to enormous volumes of data by using *crowdsourcing*, which means the labor of assigning annotations is carried out by workers recruited from online platforms such as Amazon's Mechanical Turk. Pustejovsky and Stubbs [PS12f] propose a manual to create annotated data for supervised machine learning algorithms, required for training and testing these algorithms. They divide an annotation project into stages (e.g., corpus acquisition, manual annotation, machine learning) and describe best practices for each stage. The stages are generalized by Ide [Ide17] from machine learning to collecting data for various purposes, such as for analysis, evaluation, or answering research questions regarding linguistic

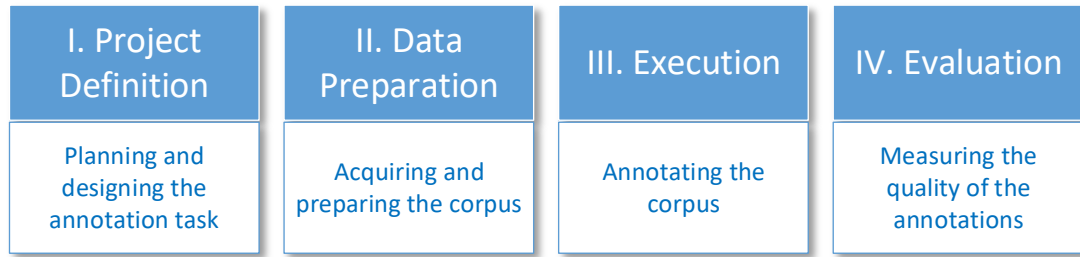


Figure 2.1: The four stages of an annotation project [SBDS14]

phenomena. The listed literature [SBDS14, PS12f, Ide17] describes best practices on corpus annotation aligned with recent research on crowdsourcing, supervised machine learning, and systematic benchmarking.

In this chapter, we summarize the background of manual corpus annotation. We divide the procedure of corpus annotation into four stages, shown in Figure 2.1. For each stage, we describe components, involved activities, and best practices. The best practices are from related literature (e.g., [Lee05, MXT06b, PS12f, SBDS14, Ide17]) and our own experiences from previous annotation projects (e.g., [ZSH<sup>+</sup>20, ZHRH20, ZABH18]).

Across the different stages, we use two annotation tasks to augment our descriptions with concrete examples. The first annotation task is sentiment classification of movie reviews. This task is about labeling a movie review’s sentiment as either *positive* or *negative*. The second task is named-entity recognition in newspaper articles. This task is about labeling mentions of *dates*, *events*, and *organizations* in newspaper articles. We refer to the first task as *sentiment analysis task* and the second as *named-entity recognition task*.

The remainder of this chapter is structured as follows: We describe the stage *project definition* in Section 2.1, *data preparation* in Section 2.2, *execution* in Section 2.3, and *evaluation* in Section 2.4. We demonstrate the application of the four-stage framework for five annotation projects in Section 2.5. The chapter is summarized in Section 2.6.

## 2.1 Project Definition

The first stage is about planning and defining an annotation project. We describe the importance of setting goals, overview different types of how labels can be assigned to text data, and describe the trade-off between accuracy and informativity when designing a new annotation task. Finally, we discuss the importance of doing background research to reuse best practices relevant to the current project.

### 2.1.1 Defining a Goal

When we launch a new annotation project, we usually have a goal in mind on how we intend to use the created annotations. Having defined a clear goal is critical, as it guides the decision-making throughout the annotation project [PS12e].

## Statement of Purpose

Defining a clear and realistic goal that can be reached given the available resources can be difficult. Pustejovsky et al. [PS12e] describe the following best practices on goal setting: The goal should be concise, comprehensive, and formulated as a *statement of purpose* using not more than two sentences. If two sentences are insufficient to express the goal, it might be too generic and requires refinement. To give an example of a statement of purpose, we consider our named-entity recognition task described in this chapter's introduction. For this task, we define the following statement of purpose:

*We want to use machine learning to automatically identify mentions of dates, events, and organizations in newspaper articles to develop an entity-specific search engine.*

This statement of purpose answers the following four key questions of an annotation project [PS12e]:

- What are the annotations used for? (search engine)
- What is the overall outcome of the annotation? (named-entity recognition)
- From where comes the corpus? (newspapers)
- How is the outcome achieved? (supervised machine learning)

Answering these key questions serves as a quick sanity check to determine whether the statement of purpose is well-defined [PS12e]. When answering these questions, it is important not to confuse the project goal with an intermediate goal: In our example, an intermediate goal is to train an automatic machine learning model, and the overall goal is to develop a search engine.

## Dividing Complex Projects

The defined goal can help identify whether a project should be carried out in the scope of one or multiple annotation efforts [PS12e]. For example, consider the goal of recognizing named-entities in German and English newspapers. Achieving this goal requires newspapers and annotators for both languages. Consequently, we might need to divide the annotation effort into two sub-efforts: the first is about annotating German newspapers and the second about English ones. Note that both sub-efforts belong to the same overall goal but require a different corpus (English and German newspaper) and probably two groups of annotators (native speakers in English and German).

Table 2.1: Common types to annotate text data

| Type                        | Example of Annotated Text  |
|-----------------------------|--|
| Document Annotation         | <i>This movie was fun and entertaining.</i> → Positive                                     |
| Text Span Annotation        | The <sup>Event</sup> <i>launch</i> of <sup>Organization</sup> <i>SpaceX</i> was postponed. |
| Linked Text Span Annotation | The <sup>Event</sup> <i>launch</i> of <sup>Organization</sup> <i>SpaceX</i> was postponed. |

### Accuracy vs. Informativity

Two properties that need consideration when formulating an annotation project’s goal is *informativity* and *accuracy* [PS12c]. Informativity is about capturing as much critical information as possible through annotation. Accuracy is the annotators’ ability to do a task sufficiently well by assigning correct annotations. There is often a trade-off between accuracy and informativity, as an increase of informativity makes a task more complex and decreases accuracy and *vice versa*. Therefore, annotation practitioners need to balance the two properties carefully.

We illustrate the trade-off between informativity and accuracy based on our named-entity recognition task. For this task, we label dates, events, and organizations in newspaper articles. Now, assume we refine the task by labeling profit and non-profit organizations. This refinement increases informativity as the annotations now capture information to distinguish between profit and non-profit organizations. However, the refined task decreases accuracy as annotators might make mistakes in labeling the new entity.

#### 2.1.2 Type of Annotation

After having formulated the project’s goal, we decide how the annotations are assigned to the text data. There are various types of text annotation, and three commonly encountered ones are *document annotation*, *text span annotation*, and *linked text span annotation* [PS12b, SBDS14]. An example of each type can be found in Table 2.1, with more details described in the following sections.

## Document Annotation

This annotation type involves the assignment of a label to an entire *document*. The concrete form of a document depends on the task and can be, e.g., a movie review, a newspaper article, or a sentence. Document annotation is typical for classification, sentiment analysis, and relevance assessment<sup>1</sup> tasks. The first example of Table 2.1 demonstrates document annotation for our task of sentiment analysis of movie reviews.

Document annotation is specified further whether one or multiple labels can be assigned per document. As an example for one label, consider our sentiment analysis task, where either the label positive or negative is assigned per movie review. As an example for multiple labels, consider the task of assigning genres to movie descriptions. For this task, multiple labels are necessary as a movie can have several genres.

## Text Span Annotation

This type involves the annotation of text parts (e.g., characters, words, phrases) within a document. Text span annotations are characteristic for named-entity recognition tasks (see the second example in Table 2.1). The annotations are associated with the text on either a *token-level* or *character-level*. Character-level means that the annotation is associated with a character position of the text. Token-level means that each document is first preprocessed using a tokenizer<sup>2</sup>, and then the annotations are associated with the tokens.

Preserving the correct associations between text span annotations and the raw text data is error-prone. When collecting the annotation on a character-level, the encoding of the text impacts the character positions. As a best practice, Pustejovsky et al. [PS12b] recommend using the universal encoding standard UTF-8. When collecting the annotations on a token-level, different tokenizers might produce a different output sequence. The recovery of the original text from a sequence of tokens is infeasible. In case such a recovery is necessary, the practitioners need to store additional information, such as the character offset of each token in the original text.

## Linked Text Span Annotation

The last discussed type is linked text span annotation to label relationships in the text. For example, in our named-entity recognition task, we can use linked text span annotations to capture relations between events and organizations (see the third example in Table 2.1). When using this annotation type, each labeled text span should have a unique identifier. The unique identifier allows storing the linked annotations using a triple  $\langle \text{ID1}, \text{ID2}, \text{Relation\_Type} \rangle$ . Note that this annotation type is based on text span annotation. Therefore, it is also affected by the text encoding and the selected tokenizer.

<sup>1</sup>Relevance assessment is a task of the IR domain to label the relevance, e.g., between a document and a search query.

<sup>2</sup>Tokenization is the process of separating a text into units based on, e.g., whitespaces and punctuation.

### 2.1.3 Background Research

Defining an annotation project from scratch is cumbersome. The practitioners need to make various critical decisions and invest resources such as time, money, and labor. Resources spent in this early stage of a project might be missing in later stages where the actual assignment of annotations is performed. To save resources, thorough background research can be beneficial.

Thorough background research is essential to identify and reuse best practices from related annotation projects. Apart from doing simple online research, practitioners might consider literature from workshops (e.g., SemEval) and conferences (e.g., ACL, LREC) related to corpus annotation [PS12e]. Even if no similar annotation project can be found, it might be beneficial to examine projects with related aspects, such as a similar goal (e.g., named-entity recognition), data source (e.g., newspapers), or annotation type (e.g., text span annotation). Overall, background research and reusing best practices increase the chance of achieving the defined project goals [PS12e].

## 2.2 Data Preparation

The next stage of an annotation project is preparing the raw text data for annotation. The raw text data organized in a structured, machine-readable format is known as the *corpus* [PS12c]. The practitioners constitute the corpus and store it, e.g., as a database or as text files. Before investing resources to create a new corpus, practitioners should check whether a suitable corpus is available for reuse. In case no suitable corpus is available, a new one can be created. The source(s) from which a new corpus is created depends on the project goal and might be, e.g., books, newspapers, websites, or research papers.

Preparing a corpus (a new or existing one) for the manual annotation process involves several design decisions. We need to decide how the annotations are stored, how large the corpus should be, and how the corpus is prepared for manual annotation. We describe best practices for these topics in the following sections.

### 2.2.1 Balance and Representativeness

The creation of a new corpus involves the sampling of text data from a target population. The population for our tasks of sentiment analysis and named-entity recognition would be movie reviews and newspaper articles, respectively. Since annotating the entire population of movie reviews or newspaper articles is infeasible, we randomly need to sample from these target populations. However, random sampling can be skewed, and therefore, we need to ensure that the sampled data is *balanced* and *representative* [Bib93, PS12e]. Representative means that the sampled text data comprises all possible text types relevant to our project goal. In our example of sampling newspaper articles, we need to acquire printed versions, online versions, newspapers from different agencies, and so on. Balanced means that a sufficient number is available for each text type. For example, when we

annotate German and English newspaper articles, we might collect an equal number of articles for both languages.

### 2.2.2 Annotation Units

Distributing the entire corpus to the annotators is often impractical. A common approach to improve the distribution and management of a corpus is to decompose it into smaller units [Ide17], which we refer to as *samples*. For example, we could generate samples for the named-entities recognition task by dividing each newspaper article into sentences, which are then distributed to individual annotators. After having all sentences annotated, we concatenate the sentences and the corresponding annotations to reconstruct the original corpus.

We differentiate between two types of decomposing a corpus into samples: a *document decomposition* and an *explicit decomposition*. Document decomposition means that each document contained in a corpus represents one sample. For example, for our task of annotating movie reviews, each movie review represents one sample of this task. On the other hand, an explicit decomposition means that the corpus is decomposed based on its linguistic structure, e.g., sentences, paragraphs, or words.

When dividing a corpus *explicitly*, the context must be preserved, as a compromised context infers with the annotators' ability to perform the task correctly. Sentences provide sufficient context for most NLP tasks [SBDS14], except for tasks that require annotations across sentences, like long-distance anaphora discovery [PCK<sup>+</sup>13]. When using software to perform the explicit decomposition (e.g., a sentence splitter), the practitioners need to be aware that this software might be error-prone and generates undesired samples.

Dividing a corpus into samples improves the managing, estimating, and planning of the available resources. For example, we can distribute the corpus to hundreds of annotators to perform a task as a collaborative effort. Furthermore, we can pay annotators per sample and estimate the average time needed to annotate one sample.

### 2.2.3 Storing

Next, we discuss techniques for storing the annotations alongside the corpus. Two techniques are commonly used [PS12b]: The first is to add the annotations directly to the text, known as *inline annotation*. The second technique is to store the annotations separately, known as *standoff annotation*. For our example of named-entity recognition, we could store inline annotations using the extensible markup language (XML) as follows:

```
<NE type="Organization">Apple</NE> was founded in <NE type="Date">1976</NE>.
```

On the other hand, for capturing the same annotation using the standoff technique, we need to separate the annotations from the corpus and store a reference linking the two. For example, we could store the two named-entities annotated in the sentence above



by storing the character offset, the character length, and the named-entity type in a comma-separated values (CSV) file as follows:

| CharOffset | CharLength | Entity       |
|------------|------------|--------------|
| 0          | 5          | Organization |
| 21         | 4          | Date         |

Storing inline and standoff annotations can be done in various ways. Consider our example task of sentiment analysis of movie reviews and assume each movie review is stored as a text file. We could directly add the sentiment label as the first line of each text file, which is a case of inline annotation. Furthermore, we could keep a separate file where we store tuples of *filename* and *sentiment label*, which is a case of standoff annotation. Finally, we could use the file structure by storing positive and negative movie reviews in two separate folders, which is another example of standoff annotation.

Although assigning annotations directly to the text seems more natural, the inline technique has two core disadvantages. First, the accompanying annotations can harm readability (see the XML tags in the example above). Second, the inline technique alters the text, and recovering the original text might be difficult. Because of these disadvantages, the best practice in storing annotations is to have a clear separation between the corpus and the annotations, making the standoff technique the preferred choice [Lee05].

### 2.2.4 Corpus Size

The amount of text data constituting a corpus is known as the *corpus size*. For document annotation tasks, we usually use the number of documents to describe the corpus size, such as the number of movie reviews, questions, or e-mails comprising the corpus. When explicitly decomposing the corpus into samples, we can describe the corpus size based on the number of samples, such as sentences, words, or paragraphs. The corpus size and the resources needed to annotate the corpus are usually correlated: We need more resources, including time, money, and annotators, to label more data volume.

The optimal corpus size depends on the project goals and is difficult to estimate. For example, we might require less volume of annotated data for creating a rule-based system than training a complex machine learning architecture. As a starting point to estimate an appropriate corpus size, related projects and their corpus sizes can be studied. As a rule of thumb, annotating larger volumes of data makes it more likely to reach the project goals [PS12e]. However, the additional effort of labeling larger volumes should not come at the cost of reduced correctness of the annotations.



## 2.3 Execution

This stage is about the process of assigning annotations to the corpus to create the *annotated corpus*. We first discuss the various components of this stage, such as the annotators, the annotation guidelines, and the annotation tool. Afterward, we describe how these components play together to create the annotated corpus.

### 2.3.1 Annotators

To create the annotated corpus, we need persons who perform the labor of assigning the annotations. These persons are known as the *annotators*. We describe how annotators are recruited, whether they need certain expertise, and how they are compensated.

#### Recruitment

A trivial approach to acquire annotators is that the practitioners themselves conduct the work. The advantage of doing the work themselves is that they know precisely how and what annotations should be assigned. By not recruiting external annotators, resources are saved that would be needed otherwise to train and prepare annotators to perform the task correctly. The disadvantage of the practitioners labeling the data is the lack of scaling, which is essential when we intend to annotate large corpora.

Another approach to recruit annotators is to consult the practitioners' immediate social circle, like friends, family, or colleagues. While these persons are often motivated to help, they might not be familiar with the task and require sufficient training to perform it accurately. However, this approach does not scale either since its limited by the size of the social circle and the time available by the individual persons.

A scalable approach for recruiting annotators is *crowdsourcing*. Crowdsourcing means that annotators are recruited from online platforms, such as Amazon's Mechanical Turk (MTurk)<sup>3</sup>. Annotators recruited from crowdsourcing platforms are known as the *crowdworkers* (also workers) and are usually available in high numbers [SBDS14]. When preparing a corpus for crowdsourcing, the annotation effort is divided into small, self-contained units. We defined these units as samples in Section 2.2.2, but in the context of crowdsourcing, they are known as *micro task* or *Human Intelligence Task* (HIT) on the Mechanical Turk platform. The term *micro* is used since tasks posted on crowdsourcing platforms are rather short, usually requiring a few seconds or minutes to be completed. After decomposing the corpus into samples, they are posted on the crowdsourcing platform, and the workers annotate them in exchange for monetary compensation. If the compensation is fair and the task appealing, sufficient workers will be motivated to participate, allowing a scalable corpus annotation.

Various aspects need to be considered when collecting annotations through crowdsourcing: First, some workers might perform the task inaccurately or try to cheat by spamming

<sup>3</sup><https://www.mturk.com>

random annotations [DDC12b]. As a best practice, a test run can be conducted to identify and recruit accurate workers [SBDS14]. Second, the effort needed to annotate samples should be evenly distributed. If the effort is unevenly distributed, workers might *cherry-pick* shorter samples and ignore longer ones [CTIB15, FSL<sup>+</sup>18]. Finally, annotations collected through crowdsourcing are often noisy with respect to their correctness [SOJN08]. A common strategy to reduce the noise is to collect redundant annotations per sample, followed by an aggregation via, e.g., a majority voting [SOJN08, SBDS14].

### Motivation and Compensation

The annotators' motivation to work on an annotation task might be *intrinsic* or *extrinsic* [DKC<sup>+</sup>18]. For fostering intrinsic motivation, the task can be designed to be fun (e.g., a game with a purpose) or serve a greater purpose (e.g., annotating data for skin cancer detection). Extrinsic motivation is fostered by fair compensation of annotators for the performed work. The usual compensation is monetary. Annotators can be paid based on their working hours or the number of labeled samples. When estimating a fair payment per labeled sample (e.g., through a test run), the practitioners need to be aware that annotators are individuals with a different pace in performing a task.

### Expertise

Some tasks require annotators with certain qualifications to be conducted accurately. These tasks are often domain-specific and require *expert annotators* with profound knowledge in the particular domain [Wal18a, ALP<sup>+</sup>16]. Expert annotators have several years of experience in the specific domain and understand its jargon and terminology. Depending on the task at hand, expert annotators can be persons such as medical practitioners (e.g., for annotating biomedical texts), lawyers (e.g., for annotating legal texts), or linguists (e.g., for annotating linguistic features).

Finding and recruiting expert annotators is usually cumbersome. They are expensive and available only in limited numbers due to being occupied with their profession. A practical approach to recruit expert annotators is collaborating with research and business partners with connections to experts. For example, when we aim to annotate biomedical texts, we might collaborate with hospitals or medical research facilities. Another possibility to recruit experts is over the internet, where suitable ones can be found through advertisement or freelancing platforms. Freelancing platforms (e.g., Upwork<sup>4</sup> and Fiverr<sup>5</sup>) allow hiring experts from a variety of domains, such as data scientists, linguists, and medical doctors.

### 2.3.2 Annotation Guidelines

When not familiar with an annotation task, annotators need to be trained to perform the task accurately. The standard approach to train annotators is to provide *annotation*

---

<sup>4</sup><https://www.upwork.com/>

<sup>5</sup><https://www.fiverr.com/>

*guidelines*. These guidelines describe how the task should be performed based on instructions and examples. The guidelines aim to align the conception between practitioners and annotators on correctly performing the task.

When crafting the guidelines, it is essential not to overwhelm the annotators with too much information. However, the guidelines should also provide enough information to prepare the annotators sufficiently well. Finding the right balance between these two properties is critical [PS12a]. The guidelines optimally describe cases on *what to do* but also on *what not to do*. Furthermore, the guidelines should (i) comprise cases that are challenging to label correctly and (ii) cover cases that are frequently encountered during an annotation task. Crafting guidelines often takes several iterations of testing and revising to obtain the final version. For some annotation projects, the annotation guidelines are publicly available so that they can be reused or adapted. Therefore, it is important to examine related projects before crafting annotation guidelines from scratch.

### 2.3.3 Annotation Tool

Assigning annotations is usually supported by a piece of software known as the *annotation tool* [FE17]. The annotation tool can be as simple as annotators using Microsoft Excel to store document IDs and corresponding labels. However, such a generic tool does not provide much support to annotators and is prone to errors (e.g., mixing up columns/rows) [Lee05]. A better alternative is to develop or reuse an annotation tool that is tailored for the task at hand. These specialized tools bridge the gap between the task, the corpus, and the annotators. When developing a new annotation tool or selecting an existing one, the practitioners should make various considerations, such as:

- Is the tool easy and convenient to use?
- How can annotators access the tool?
- Does the tool offer sufficient functionality to assign valid annotations?
- Does the tool forbid invalid annotations?
- How does the tool represent the corpus?

The recent trend in developing annotation tools is to implement a server-side back end accessible through a web-based interface. Having a web-based interface allows annotators to access the tool through their web browser without requiring a local software installation. An online annotation tool enables seamless information exchange between the annotators and the practitioners: The annotators label the data, which is then stored on the server (e.g., in a database) and finally downloaded by the practitioners.

Web-based annotation tools are the standard for crowdsourcing-based annotation. Most crowdsourcing platforms offer predefined templates for various annotation types, such

as document annotation and text span annotation. After selecting a template, the practitioners decompose the corpus into samples and upload them to a crowdsourcing platform of choice. The labeled samples are downloaded through a platform-specific management tool, which also offers functionality to monitor the crowdsourcing process continuously for quality control [SBDS14].

Developing a new annotation tool from scratch is a time-intensive procedure, especially when developing a tool with an intrinsic incentive (e.g., a game with a purpose) [SBDS14]. Rather than developing a new tool from scratch, existing annotation tools are available for many different tasks. Popular open-source tools for NLP-related annotation tasks are BRAT [SPT<sup>+</sup>12], GATE [CMB11], and Doccano<sup>6</sup>. Templates offered by crowdsourcing platforms are also freely available, and they can be customized with basic HTML and JavaScript skills. Commercial software is available by, e.g., Prodigy<sup>7</sup>, Tagtog<sup>8</sup>, and Labelbox<sup>9</sup>.

### 2.3.4 Annotation Process

We described various components of an annotation project so far, such as the statement of purpose, the corpus, the annotators, the guidelines, and the annotation tool. Next, we describe how these components play together to create the annotated corpus in a process referred to as the *annotation process*. The typical annotation process consists of the following steps [PS12a]: First, the annotation guidelines are crafted aligned with the defined statement of purpose. Then, the guidelines and the corpus are distributed to the annotators, usually via the annotation tool. Afterward, the annotations are assigned by the annotators. Finally, the annotated data is obtained and evaluated by the practitioners.

The annotation process is often cyclic, requiring various iterations of testing and adapting. Pustejovsky et al. [PS12a] describe the process as the Model-Annotate-Model-Annotate cycle (MAMA cycle), illustrated in Figure 2.2. Whether another iteration of revising is necessary is determined through small-scale test runs using only a fraction of the corpus and the annotators. These test runs can help identify and refine weak components of an annotation project. The practitioners decide whether another iteration is necessary by evaluating the quality of the most recently collected annotations. We describe how the quality of annotations is evaluated as part of the next stage.

## 2.4 Evaluation

The quality and reliability of an annotated corpus can be measured using the inter-annotator agreement [McH12, FE17]. The inter-annotator agreement measures the

---

<sup>6</sup><https://github.com/doccano/doccano>

<sup>7</sup><https://prodi.gy/>

<sup>8</sup><https://www.tagtog.net/>

<sup>9</sup><https://labelbox.com/>

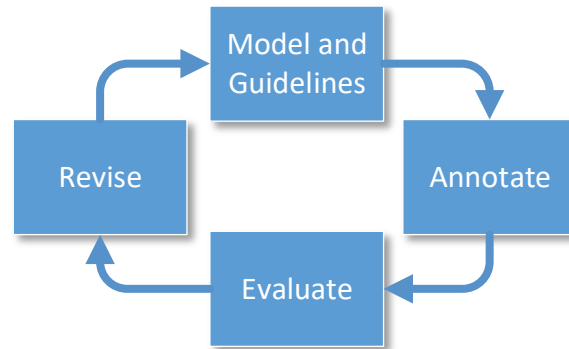


Figure 2.2: The Model-Annotate-Model-Annotate (MAMA) cycle proposed in [PS12a]

agreement between several annotators for labeling a shared set of samples [McH12]. A high inter-annotator agreement indicates a similar conception between annotators of how the data should be labeled. On the other hand, a low inter-annotator agreement indicates frequent disagreement between annotators on what labels should be assigned. The agreement between annotators is an indirect measure of the reliability of an annotated corpus [McH12]: If not even human annotators can agree on how the task should be performed, then an automatic solution usually also fails to do so.

We first define how the inter-annotator agreement is computed. Afterward, we describe how the computed agreement is interpreted, and finally, we discuss commonly encountered challenges causing disagreement.

### 2.4.1 Cohen’s Kappa Agreement

A naive approach to compute the agreement between two annotators is to compute the relative agreement, defined as:

$$p_0 = \frac{\#\text{Agreeing}}{\#\text{Agreeing} + \#\text{Disagreeing}}, \quad (2.1)$$

where  $\#\text{Agreeing}$  is the number of samples for that both annotators agree, and  $\#\text{Disagreeing}$  is the number of samples where they disagree.

However, the relative agreement does not consider the probability of annotators agreeing by chance. A robust metric regarding agreement by chance is the Cohen’s Kappa statistic (also referred to as the Kappa agreement). The Kappa statistic is a standard metric to compute the inter-annotator agreement between two annotators [McH12] and is defined as:

$$\kappa = \frac{p_0 - p_c}{1 - p_c}, \quad (2.2)$$

where  $p_0$  is the relative agreement, and  $p_c$  is the probability of agreement by chance. The probability of agreement by chance for labeling  $N$  samples when  $k$  possible labels can be assigned per sample is defined as

$$p_c = \frac{1}{N^2} \sum_k n_{k1} \times n_{k2}, \tag{2.3}$$

where  $n_{ki}$  is the frequency of annotator  $i$  assigning the label  $k$ .

We illustrate how the Cohen’s Kappa agreement is computed and its difference to the relative agreement based on an example. Consider our sentiment analysis task of labeling movie reviews as either *positive* or *negative*. Three movie reviews are labeled by two annotators as follows:

| Movie Review | Annotator1 | Annotator2 |
|--------------|------------|------------|
| Review #1    | positive   | positive   |
| Review #2    | negative   | negative   |
| Review #3    | negative   | positive   |

Both annotators agree for the first two reviews and disagree for the third one, resulting in a relative agreement of  $p_0 = \frac{2}{2+1} \approx 0.66$ . The probability of labeling a review as positive is  $\frac{1}{3}$  for the first annotator and  $\frac{2}{3}$  for the second, resulting in a combined probability of labeling a review as positive of  $\frac{1}{3} \times \frac{2}{3} \approx 0.22$ . The probability of labeling a review as negative is  $\frac{2}{3}$  for the first annotator and  $\frac{1}{3}$  for the second, resulting in a combined probability of labeling a review as negative of  $\frac{2}{3} \times \frac{1}{3} \approx 0.22$ . By summing up these two probabilities, we obtain the probability of the annotators agreeing by chance of  $p_c = 0.22 + 0.22 = 0.44$ . We now calculate the Kappa score  $\kappa = \frac{p_0 - p_c}{1 - p_c} = \frac{0.66 - 0.44}{1 - 0.44} \approx 0.39$ . Notice that the Kappa score is substantially lower than the relative agreement of  $p_0 \approx 0.66$  because of the normalization of agreeing by chance.

### 2.4.2 Kappa Agreement for Text Span Annotation

The previous example showed how the Kappa statistic is computed for document annotation tasks. However, we require a slightly different approach for text span annotation, as often an unequal number of annotations are assigned, depending on the text parts labeled by the annotators.

We show how the Kappa agreement is computed for text span annotations based on an example: Assume we collect token-level text span annotations (see Section 2.1.2) for our task of named-entity recognition in newspaper articles. A sentence is labeled by two annotators as follows:

Annotator 1 = The  $\overbrace{\text{launch}}^{\text{Event}}$  of  $\overbrace{\text{SpaceX}}^{\text{Organization}}$  was postponed .  
 Annotator 2 =  $\overbrace{\text{The launch of}}^{\text{Event}}$   $\overbrace{\text{SpaceX}}^{\text{Organization}}$  was postponed .

Next, we encode these annotations into two sequences of equal length. The sequences are derived from mapping each token to a label using the following dictionary:

- *None*: The token is not annotated
- *Event*: The token is annotated as an event
- *Organization*: The token is annotated as an organization

After mapping each token using the defined dictionary, we obtain the following two sequences:

| Token            | Annotator1   | Annotator2   |
|------------------|--------------|--------------|
| <i>The</i>       | None         | Event        |
| <i>launch</i>    | Event        | Event        |
| <i>of</i>        | None         | Event        |
| <i>SpaceX</i>    | Organization | Organization |
| <i>was</i>       | None         | None         |
| <i>postponed</i> | None         | None         |
| .                | None         | None         |

These sequences are now used with the Kappa statistic from Equation 2.2 to compute the inter-annotator agreement of  $\kappa \approx 0.53$ .

The example demonstrated how the Kappa statistic is computed for token-level text span annotations. For character-level annotation, the sequences are derived from mapping each character to a label instead of tokens. In case annotations overlap (e.g., a token might have been annotated as an *event* and *organization*), a new dictionary entry can be added, such as *Event-Organization*.

### 2.4.3 Interpretation of Kappa Agreement

The Cohen’s Kappa score  $\kappa$  ranges from -1 (full disagreement) to +1 (full agreement). Related work aims to interpret the Kappa score based on strictly defined intervals, as shown in Table 2.2. Interpreting the Kappa score based on fixed intervals is a starting point, but also other aspects of a task (e.g., difficulty, domain-specificity) must be considered for interpretation [PS12a]. For example, consider our sentiment analysis task, where we label movie reviews as either positive or negative. The Kappa agreement will likely decrease for this task if we change its complexity by annotating the sentiment using integer values from 1 (very negative) to 10 (very positive).

### 2.4.4 Challenges of Agreement

Aiming for high Kappa agreement is essential, as it indicates the quality and reproducibility of annotations [PS12a]. Obtaining a high Kappa agreement can be challenging, depending on various factors of an annotation project. Some factors relate to proper decision-making, such as recruiting qualified annotators, designing comprehensive annotation guidelines, and developing a user-friendly annotation tool. Other factors are specific to certain tasks and domains [PS12a], described in the following sections.

#### Subjectivity

Agreement depends on a common conception between annotators and practitioners on how a task should be performed. The conception is influenced by the annotators’ subjectivity, shaped by their education, background, and experiences from previous annotation tasks [GCC12, LBRD<sup>+</sup>20]. As an example of subjectivity, consider the task of genre assignment to movie descriptions. One annotator might find a movie funny and assigns the genre *comedy*, while another annotator might not share the same opinion. Tasks more affected by subjectivity are those that ask for opinions



Table 2.2: Interpretation intervals for the Cohen’s Kappa score  $\kappa$ 

| (a) Proposed by McHugh [McH12] |                | (b) Proposed by Landis and Koch [LK77] |                |
|--------------------------------|----------------|--|----------------|
| Cohen’s Kappa ( $\kappa$ )     | Interpretation | Cohen’s Kappa ( $\kappa$ )             | Interpretation |
| < 0.20                         | None           | < 0.00                                 | Poor           |
| 0.21-0.39                      | Minimal        | 0.00-0.20                              | Slight         |
| 0.40-0.59                      | Weak           | 0.21-0.40                              | Fair           |
| 0.60-0.79                      | Moderate       | 0.41-0.60                              | Moderate       |
| 0.80-0.90                      | Strong         | 0.61-0.80                              | Substantial    |
| 0.90-1.00                      | Almost Perfect | 0.81-1.00                              | Almost Perfect |

from annotators, such as labeling sentiment or relevance (e.g., between a search query and a document) [MRS08].

The subjectivity of a task can be tuned via specification and generalization [PS12d]. For example, consider our task of sentiment analysis of movie reviews. We can increase this task’s specification by annotating fine-grained sentiment labels from 1 (very negative) to 10 (very positive). While this increases specificity, the task is now more subjective, as annotators might disagree, e.g., whether a movie should be labeled as 8 or 9. We can reduce the task’s subjectivity by generalizing to the original binary differentiation between *positive* and *negative*. Although this generalization reduces subjectivity, the binary task might interfere with our project goal, which could be a fine-grained sentiment annotation using a 10-class system.

### Consistency

Next, we discuss the issue of annotators being inconsistent with their annotations. Inconsistency is a characteristic problem of text span annotation (e.g., named-entity recognition), as these require precise labeling of tokens or characters within a text. For example, consider our task of named-entity recognition in newspapers and the following three annotations:

$$\begin{array}{c} \text{Date} \\ \overbrace{\text{January the 2}^{\text{nd}} \text{ at 10:14 PM}} \\ \text{Date} \quad \text{Date} \quad \text{Date} \\ \underbrace{\text{January}} \text{ the } \underbrace{\text{2}^{\text{nd}}} \text{ at } \underbrace{\text{10:14 PM}} \\ \text{Date} \quad \text{Date} \\ \underbrace{\text{January}} \text{ the } \underbrace{\text{2}^{\text{nd}}} \text{ at 10:14 PM} \end{array}$$

These three phrases were inconsistently annotated: the first annotator labeled the entire phrase, the second labeled each date individually, and the third left out the time. Consistency issues like the presented one are usually addressed via the annotation guidelines. Before being addressed in the guidelines, we need to identify potential consistency issues. Test runs and studying related tasks might provide sufficient information to forecast consistency issues. However, identifying and addressing all possible consistency issues is often infeasible and would lead to a lengthy guideline document, potentially overwhelming the annotators.



## Domain-Specificity

As a final challenge, we discuss domain-specific tasks for which obtaining high-quality annotations is usually difficult. Domains characteristic for difficult annotation tasks are the legal, patent, and medical domain [ALP<sup>+</sup>16]. These domains have dense jargon, which is difficult to understand and annotate by non-experts. However, even by recruiting experts, we cannot expect a full agreement between the annotators. Each expert has a unique background, with different specialization to conduct a task accurately. Alone the biomedical domain has more than 135 medical specialties<sup>10</sup>, such as mental, dental, or digestive. The same applies to the legal domain with more than 20 specifications<sup>11</sup>, such as building law, tenancy law, and family law. Even combinations of several domains are possible such as tasks concerning medical law. Because of this versatility, domain-specific tasks pose a substantial challenge in achieving high annotator agreements, and even more so when employing non-expert annotators [NLP<sup>+</sup>18, ZAH<sup>+</sup>18].

## 2.5 Examples of Annotation Projects

We now give examples of annotation projects following the four stages of project definition, data preparation, execution, and evaluation. We conduct five projects about the following tasks:

- **Query-document relevance assessment:** The aim of this task is to label the relevance of a document with respect to a search query. Annotators need to assign one out of the following four classes per query-document pair: perfectly relevant (i.e., *perfect*), partially relevant (*partial*), topic relevant (*topic*), and not relevant at all (*wrong*).
- **Query-document text span labeling:** Given a search query and a document, the aim of this task is to highlight text spans in the document that are relevant with respect to the query's information need. We conduct this task combined with the query-document relevance assessment task.
- **Biomedical named-entity annotation:** For this task, the annotators highlight text spans in clinical study reports for the named-entities: *Participant* (e.g., patients with headache), *Intervention* (e.g., Thomapyrin), and *Comparison* (e.g., placebo).
- **Clinical study polarity analysis:** The annotators label the polarity of clinical study reports. The polarity indicates whether the outcome of a conducted study is *positive*, *negative*, or *neutral*.
- **Disease-symptom relevance assessment:** For disease-symptom pairs, the aim is to label whether the symptom is a primary symptom with respect to the disease. Primary symptoms are the key symptoms for disease diagnosis.

This section demonstrates the wide variety of IR and NLP annotation tasks that fit into the four-stage framework. For each stage and project, we describe the annotation-related design decisions. For the evaluation stage, we analyze the inter-annotator agreement and show the challenges of subjectivity, consistency, and domain-specificity. An overview of the five projects is available in Table 2.3. The table summarizes characteristics of the tasks, corpora, annotators, annotation guidelines, and annotation tools. We describe further details on each project in the following sections.

<sup>10</sup>Information obtained from the Association of American Medical Colleges (AAMC), available at <https://www.aamc.org/cim/explore-options/specialty-profiles>

<sup>11</sup>Information obtained from Germany's Bundesrechtsanwaltskammer, available at <https://brak.de/fuer-journalisten/zahlen-zur-anwaltschaft/>

## 2. BACKGROUND ON MANUAL CORPUS ANNOTATION

Table 2.3: Overview of the characteristics of the five annotation projects

| Task                                 | Annotation Type      | Domain      | Corpus  | Corpus Size     | Number of Annotators | Annotator Profession      | Annotation Guidelines     | Annotator Recruitment    | Annotator Motivation | Annotation tool |
|--------------------------------------|----------------------|-------------|---|-----------------|----------------------|---------------------------|---------------------------|--------------------------|----------------------|-----------------|
| Query-document relevance assessment  | Document Annotation  | General     | Queries and documents of Bing's search Engine                     | 24,199 pairs    | 87                   | Computer science students | Examples and instructions | Students from IR lecture | Part of exercise     | Web-based       |
| Query-document text span labeling    | Text Span Annotation | General     | Queries and documents of Bing's search Engine                     | 24,199 pairs    | 87                   | Computer science students | Examples and instructions | Students from IR lecture | Part of exercise     | Web-based       |
| Biomedical named-entity annotation   | Text Span Annotation | Bio-medical | Abstracts of clinical trial reports                               | 1,416 abstracts | 5                    | Medical librarians        | Examples and instructions | Research collaboration   | Monetary             | Web-based       |
| Clinical study polarity analysis     | Document Annotation  | Bio-medical | Abstracts of clinical trial reports                               | 1,147 abstracts | 5                    | Medical librarians        | Examples and instructions | Research collaboration   | Monetary             | Web-based       |
| Disease-symptom relevance assessment | Document Annotation  | Bio-medical | Disease-symptom pairs from medical textbooks and online resources | 232 pairs       | 3                    | Medical doctors           | Examples and instructions | Research collaboration   | Monetary             | Pen and paper   |

### 2.5.1 Project Definition

#### Query-Document Relevance Assessment

The first annotation task is the relevance assessment of a document with respect to a search query. This task is important to create new test collections for evaluating IR systems. The manual assessment of a query-document pair is usually conducted on a binary level by assigning either the label *relevant* or *irrelevant* [VH02, VH03]. However, a binary differentiation is often not sufficient to describe the relevance between a query and a document, e.g., consider a case where a document only partially answers a query's information need. Therefore, we label query-document pairs using the following four classes:

- **Wrong:** The document does not answer the query's information need.
- **Topic:** The document does not answer the query but is on the same topic.
- **Partial:** The document answers the query partially.
- **Perfect:** The document answers the query fully.

#### Query-Document Text Span Labeling

This task is about labeling text spans in documents that are relevant with respect to a search query. Documents are usually long, and not the entire text might be relevant with respect to a search query. Therefore, a research direction of IR is to retrieve information from documents on a more fine-grained level, such as passages [BCC<sup>+</sup>16] or sentences [BAC07]. The evaluation of such fine-grained retrieval systems depends on the availability of suitable test collection. We create such a test collection through manual annotation.

#### Biomedical Named-Entity Annotation

For this task, we acquire named-entity annotations for Evidence-Based Medicine. Evidence-Based Medicine is the practice of decision-making based on the best evidence available [SRG<sup>+</sup>96]. A commonly consulted resource for conducting Evidence-Based Medicine is clinical trial reports. These reports are published studies in which the effect of a new treatment is tested in patients with a certain medical condition. The patients are usually divided into a *treatment group*, receiving a new treatment method, and a *control group*, receiving traditional treatment. Since over 33,400 clinical trial reports were published only in 2019<sup>12</sup>, finding the best available evidence requires a tremendous effort from medical practitioners. For enhancing the search procedure, text mining approaches were developed (e.g., [Aro01, GSR18]) to automatically extract information from clinical trial reports. These text mining approaches are trained and evaluated using annotated corpora. We create such a corpus by labeling the following named-entities in clinical trial reports:

- the medical condition and characteristics of the patients (referred to as *Participant*)
- the treatment administered to the treatment group (referred to as *Intervention*)
- the treatment administered to the control group (referred to as *Comparison*)

<sup>12</sup>Source is the medical publication database PubMed: <https://pubmed.ncbi.nlm.nih.gov/>

Table 2.4: Sentences with annotated polarity classes of *positive*, *negative*, and *neutral*.

| Polarity | Sentence   |
|----------|--|
| Positive | Venous microangiopathy was improved by the treatment with Venoruton.                         |
| Negative | Ginger is not a clinically relevant antiemetic in the PONV setting.                          |
| Neutral  | There was no significant difference between the active treatments on either TI or UPDRS III. |

We label the named-entities Participant, Intervention, and Comparison (PIC) within the title and the abstract of clinical trial reports. The annotation of titles and abstracts instead of the full-text publication is the standard practice for clinical reports [KMCY11] since the title and the abstract usually contain a condensed description of all the PIC information.

We provide additional guidance to annotators by indicating mentions of drugs, diseases, and persons in the texts of the clinical trial reports. These three semantic labels are often related to the PIC information. Mentions of diseases and persons are usually related to the Participants of a study. Mentions of drugs are usually related to the Intervention or Comparison. The semantic labels for drugs and diseases are automatically generated using GATE’s BioYodie pipeline [GSR18], a tool for named-entity recognition in medical documents. To label person mentions, we create a static lookup list consisting of 44 person keywords (e.g., *patients*, *seniors*, *children*). Note that the semantic labels cannot be used as a replacement for extracting structured PIC information since PIC entities are usually longer phrases providing additional context to diseases, drugs, and persons (e.g., consider the Participant phrase *Japanese type 2 diabetic patients*) [HLD06].

### Clinical Study Polarity Analysis

This task is about classifying the polarity of outcomes reported in clinical trial reports. The polarity indicates whether the outcome of a tested research hypothesis is positive, negative, or neutral. The polarity depends on the evaluated parameters, which could be, for instance, safety, efficacy, or drug tolerance. The automatic extraction of the polarity of a trial report contributes to Evidence-Based Medicine by allowing practitioners to formulate polarity-specific search queries [NZLH05]. For example, a practitioner might search for drugs with high efficacy for treating diabetes. Annotated datasets are required to automate the polarity classification of clinical trial reports based on supervised learning. However, there is limited research on the topic, with a lack of publicly available datasets [NZLH05, DD15]. Therefore, to create such a resource, we perform the annotation task of polarity analysis of clinical trial reports.

We assign the polarity to sentences appearing in clinical trial reports. Each sentence of an abstract is labeled with one of the following classes: *positive*, *negative*, or *neutral*. We show examples of each polarity class in Table 2.4. Note that sentences that do not contain any polarity-specific information are labeled as *neutral*. We assign labels to sentences rather than the entire abstract since a clinical trial report might contain a mixed polarity. For example, one sentence might mention the high efficacy of a new treatment method, whereas another sentence reveals that the new treatment method has serious side effects. Furthermore, the annotation of individual sentences allows computing an overall polarity for a trial report, e.g., by a majority voting.

## Disease-Symptom Relevance Assessment

In this task, we label the relevance of symptoms with respect to diseases. Disease-symptom knowledge bases are the foundation for many medical tasks, such as computer-assisted diagnosis [NFFZ17] or the analysis of unexpected relations between diseases and symptoms [ZMBS14, dVGS<sup>+</sup>18]. Most knowledge bases only capture a binary relationship between diseases and symptoms, neglecting the degree of the importance between a symptom and a disease. For example, abdominal pain and nausea are both symptoms of appendicitis, but while abdominal pain is a key differentiating factor, nausea does little to distinguish appendicitis from other digestive diseases. While several disease-symptom extraction methods have been proposed that retrieve a ranked list of symptoms for a disease [ZMBS14, SLK<sup>+</sup>19, MBC14, XSM<sup>+</sup>18], no dataset is available to evaluate the performance of such methods systematically [SLZ<sup>+</sup>19].

Therefore, we create a new annotated dataset for the task of disease-symptom relevance assessment. We label disease-symptom pairs using graded judgments [Kek05] by differentiating between *relevant symptoms* (graded as 1) and *primary symptoms* (graded as 2). Primary symptoms—also called cardinal symptoms—are the symptoms that guide physicians in disease diagnosis. The consideration of graded judgments allows for the first time to measure the importance of different symptoms with grade-based metrics, such as nDCG [JK02].

### 2.5.2 Data Preparation

We now describe the corpora that are annotated in the individual tasks. We obtain the underlying data for each task from publicly available resources. For the tasks relevance assessment and text span labeling of query-document pairs, we use the TREC Deep Learning Track from 2019 [CMY<sup>+</sup>20] as the data source. The data used in the TREC Deep Learning Track was derived from the MS MARCO collection [BCC<sup>+</sup>16], containing websites and user search queries of Microsoft’s Bing search engine. Since the website texts are usually long and unequal in length, we divide them into documents of similar length as follows: First, the website text is split into sentences<sup>13</sup>. Then, the sentences are sequentially concatenated into documents so that a maximum length of 130 words is not exceeded per document. Our sampling approach generates semantically coherent documents of similar length with approximately 120-130 words each. In total, we obtain 24,199 query-document pairs, which we use for the annotation tasks relevance assessment and text span labeling of query-document pairs.

The data for the tasks biomedical polarity analysis and named-entity annotation are sampled from PubMed, a database containing entries for more than 30 million biomedical publications<sup>14</sup>. PubMed is a frequently consulted resource by IR and NLP researchers since for each publication, the title, abstract, and various other meta-data are freely available [NLP<sup>+</sup>18]. Publications can be accessed via an application programming interface (API) or through the website. When retrieving publications from PubMed, the user can define various filtering criteria, such as filtering based on publication year or publication type. We retrieve clinical trial reports by filtering for the publication type of *Clinical Trial* and *Randomized Controlled Trial*<sup>15</sup> appearing from 2006 to 2018. In total, we obtain 278,112 trial reports. From these, we randomly sample and annotate 1,147 reports for the polarity analysis task and 1,416 reports for the PIC named-entity annotation

<sup>13</sup>We use the BlingFire library to segment sentences.

<sup>14</sup>As of November 2020 stated on <https://pubmed.ncbi.nlm.nih.gov>

<sup>15</sup>A randomized controlled trial is a clinical trial with a specific study design in which patients are randomly assigned into different treatment groups.

task. We annotate the polarity over the sentences of the trial reports and the PIC entities over the tokens. For tokenization and sentence segmentation, we use the CoreNLP library [MSB<sup>+</sup>14].

The data samples for the task of labeling primary symptoms of disease-symptom pairs are prepared by two physicians of a Viennese hospital. The two physicians collect disease-symptom pairs (e.g., *appendicitis-nausea*) in a collaborative effort from high-quality sources, including medical textbooks and an online information service<sup>16</sup> that is curated by medical experts. In total, they obtain 232 disease-symptom pairs for 20 diseases (i.e., an average of  $\approx 11.6$  symptoms per disease). As a final step in preparing the data samples, we map each disease and symptom to the Unified Medical Language System (UMLS) vocabulary. The UMLS vocabulary contains unique identifiers for over four million medical concepts such as diseases, symptoms, or body parts. The mapping of the diseases and symptoms to a medical vocabulary allows unambiguous identification of the medical terms independent of language or spelling. We use the UMLS vocabulary since it is a compendium of over 200 vocabularies (e.g., ICD-10, MeSH, SNOMED-CT) cross-linked with each other. The cross-linking allows converting the disease/symptoms entries from one medical vocabulary to another and makes our dataset compatible with all of the 200 different vocabularies.

### 2.5.3 Execution

#### Annotators and Annotation Guidelines

To label the disease-symptom pairs, we recruit three physicians who work in the emergency and intensive care department of a Viennese hospital. The annotators work in the same hospital as the two other physicians who prepared the disease-symptom pairs. They have a degree in general medicine and diagnose illnesses on a daily basis. Their medical experience and education are essential for distinguishing primary and non-primary symptoms accurately.

The annotators of the task relevance assessment and text span labeling of query-document pairs are 87 computer science students. The students were recruited in the scope of the *Advanced Information Retrieval* course held in the winter term of 2020 at the Technical University of Vienna (TU Wien). We prepared the students for the annotation task by giving an introductory lecture on manual data acquisition for IR-related tasks. This lecture taught the students about the query-document annotation procedure and its importance for evaluation and supervised training.

For the labeling of named-entities and polarity in clinical trial reports, we employ five medical-domain experts. They work as librarians in medical facilities in the United Kingdom and are familiar with the jargon and terminology appearing in biomedical publications. The annotators are English native speakers, which is the same language as the clinical trial reports.

We prepare the annotators of each task with annotation guidelines consisting of instructions and examples. These instructions and examples aim to align the conception on how the task should be performed. Annotators with a similar conception usually agree more frequently, increasing the inter-annotator agreement and the created dataset's quality. When creating the guidelines, we balance conciseness and comprehensiveness. Overall, the guidelines of each task are relatively short, with about 2-5 pages. Only for the PIC labeling task, we provide a more extensive guideline document with 12 pages because this task is slightly more complex, requiring specific instructions and examples for each PIC entity. The annotation guidelines of the conducted tasks are available in Appendix A.

---

<sup>16</sup>The website <https://www.netdoktor.at>, certificated by the Health on the Net Foundation

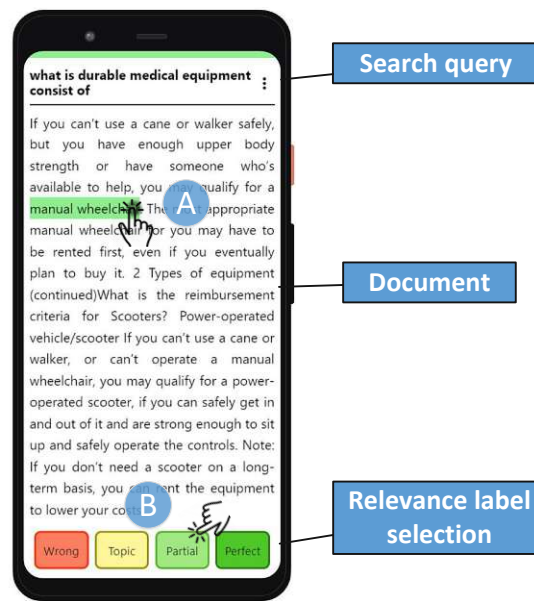


Figure 2.3: User interface of the tasks query-document relevance assessment and text span labeling. For text span labeling, annotators mark the characters within the document that answer the query’s information need (shown in A). For relevance assessment, annotators rate the document’s relevance by selecting one of the four classes (shown in B).

### Annotation Interfaces

We now shift to describing the annotation tools used in the different tasks. We develop annotation tools tailored for each task. The web-based annotation interface for collecting text span annotations for query-document pairs is illustrated in Figure 2.3. The interface combines the two tasks of query-document text span labeling and relevance assessment. The interface shows a query-document pair for which the annotator selects one of the four relevance classes and highlights relevant text spans. The interface is optimized for display on desktop computers and mobile devices.

Another web-based interface is developed for PIC labeling, shown in Figure 2.4. The interface shows the title and the abstract of a clinical trial report in which the annotators assign labels for Participant, Intervention, and Outcome. Note that we implemented a sentence navigation, allowing annotators to go through each sentence for assigning labels. We use a sentence-based navigation rather than a free token selection within the entire text for the following two reasons: First, by navigating through sentences, annotators examine each sentence at least once and might conduct the task more carefully. Second, we can give guidance (see Section 2.5.1) for each sentence by indicating the semantic labels for drugs, diseases, and persons (as shown in (C) of Figure 2.4).

The annotation interface for the polarity analysis task is shown in Figure 2.5. The interface shows the abstract of a clinical trial report. The annotator can select each sentence of a trial report to assign a polarity label of either positive, neutral, or negative. Sentences that were not explicitly selected and labeled by the annotator are considered *neutral* per default.



## 2. BACKGROUND ON MANUAL CORPUS ANNOTATION

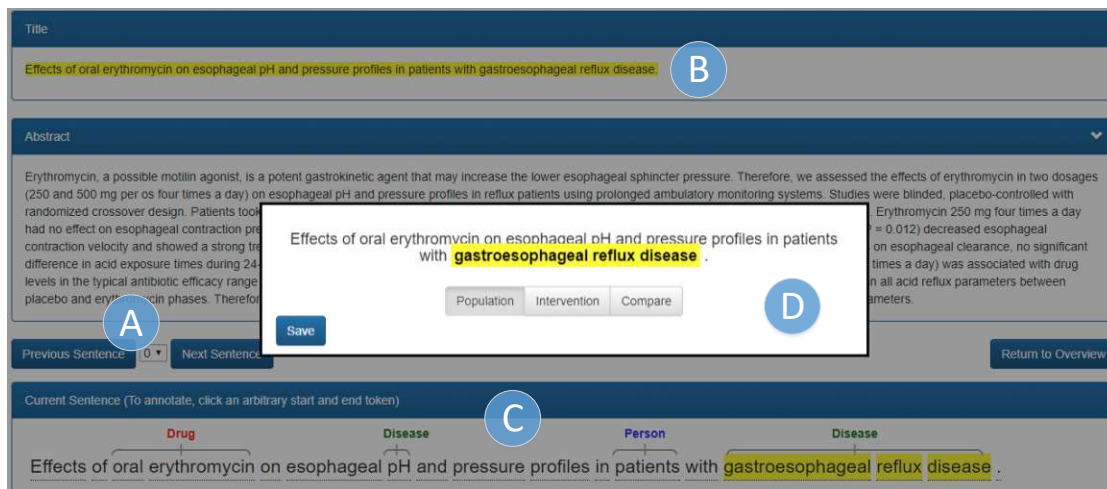


Figure 2.4: User interface of the PIC annotation task with following components: (A) sentence navigation, (B) active sentence (yellow background), (C) active sentence split into tokens, and (D) selection of the PIC label. The PIC label is assigned through a pop-up window shown as soon as a token range is selected within the active sentence.

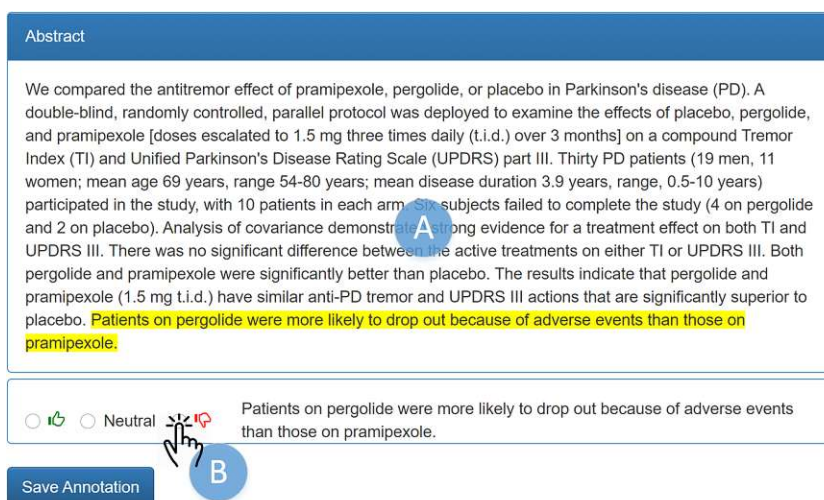


Figure 2.5: User interface of the task polarity analysis of clinical studies. The full abstract (shown in A) is presented to annotators who label the polarity of individual sentences. The currently selected sentence, highlighted by a yellow background, can be labeled as positive, neutral, or negative (shown in B).



| Disease (GER) | Disease (ENG) | Symptom (GER)                          | Symptom (ENG)        | Primary Symptom          |
|---------------|---------------|--|----------------------|--------------------------|
| Parodontitis  | Periodontitis | Eiterentwicklung                       | Pus                  | <input type="checkbox"/> |
|               |               | Entzündung des Zahnfleisches           | Gingivitis           | <input type="checkbox"/> |
|               |               | Lockere Zähne                          | loose tooth or teeth | <input type="checkbox"/> |
|               |               | Zahnfleischbluten                      | Gingival Hemorrhage  | <input type="checkbox"/> |
|               |               | Angeschwollenes Zahnfleisch            | Dental swelling      | <input type="checkbox"/> |
|               |               | Zahnfleischschwund (Gingiva-Rezession) | Gingival Recession   | <input type="checkbox"/> |
|               |               | Ungewöhnlicher Mundgeruch              | Halitosis            | <input type="checkbox"/> |

Figure 2.6: Annotations of primary symptoms were collected by providing sheets of paper with disease-symptom pairs to the annotators. The figure shows disease-symptom pairs for the disease periodontitis. Notice that we included the German names for diseases and symptoms since the annotators are German native speakers.

For the task of labeling primary symptoms, we use a slightly different annotation tool. We conduct the task by providing sheets of paper with disease-symptom pairs to the annotators. We show an example of our paper-based version for labeling disease-symptom pairs of periodontitis<sup>17</sup> in Figure 2.6. We opted for this analog task design based on the annotators’ feedback. The annotators—who work as medical practitioners in Viennese hospitals—found a paper-based version convenient to work on during short breaks in the hospital. We manually transferred the filled paper sheets into a machine-readable, comma-separated value list. We checked the resulting list three times to make sure that no mistake occurred during the manual transfer.

### 2.5.4 Evaluation

We now arrive at the last stage: the evaluation of the collected annotations. We give an overview of the collected annotations of each task in Table 2.5. Each data sample (e.g., query-document pair, clinical trial report, disease-symptom pair) is labeled by multiple annotators. For the two query-document labeling tasks and the disease-symptom labeling task, each pair is labeled by three annotators on average. For the task biomedical named-entity annotation and polarity analysis, each clinical trial report is labeled by two annotators on average. By having samples annotated by multiple workers, the quality of a dataset can be improved by aggregating redundant labels into a meta-label of higher quality (e.g., via a majority voting). Furthermore, the collection of redundant labels allows us to analyze the inter-annotator agreement.

#### Inter-Annotator Agreement

We compute for each task the inter-annotator agreement using the Cohen’s Kappa statistic (described in Section 2.4). We report the Cohen’s Kappa statistic between annotators of each task in Figure 2.7. For the three biomedical tasks, we observe average agreements ranging from 0.52 to 0.66. Lower Kappa scores are found for the task of query-document relevance assessment and text span labeling. For the two query-document labeling tasks, notice a much higher variance of Kappa scores compared to the three biomedical tasks. The high variance is caused by the high number of 87 annotators who participated compared to the three biomedical tasks where only 3-5 annotators participated. To better understand the causes of disagreement of each task, we analyze task-specific challenges next.

<sup>17</sup>A dental disease where the gum that surrounds the teeth retreats

Table 2.5: Statistics of the five annotated datasets

| Task                                 | Number of Annotators | Corpus Size     | Annotations per Sample |
|--------------------------------------|----------------------|-----------------|------------------------|
| Query-document relevance assessment  | 87                   | 24,199 pairs    | 3                      |
| Query-document text span labeling    | 87                   | 24,199 pairs    | 3                      |
| Biomedical named-entity annotation   | 5                    | 1,416 abstracts | 2                      |
| Clinical study polarity analysis     | 5                    | 1,147 abstracts | 2                      |
| Disease-symptom relevance assessment | 3                    | 232 pairs       | 3                      |

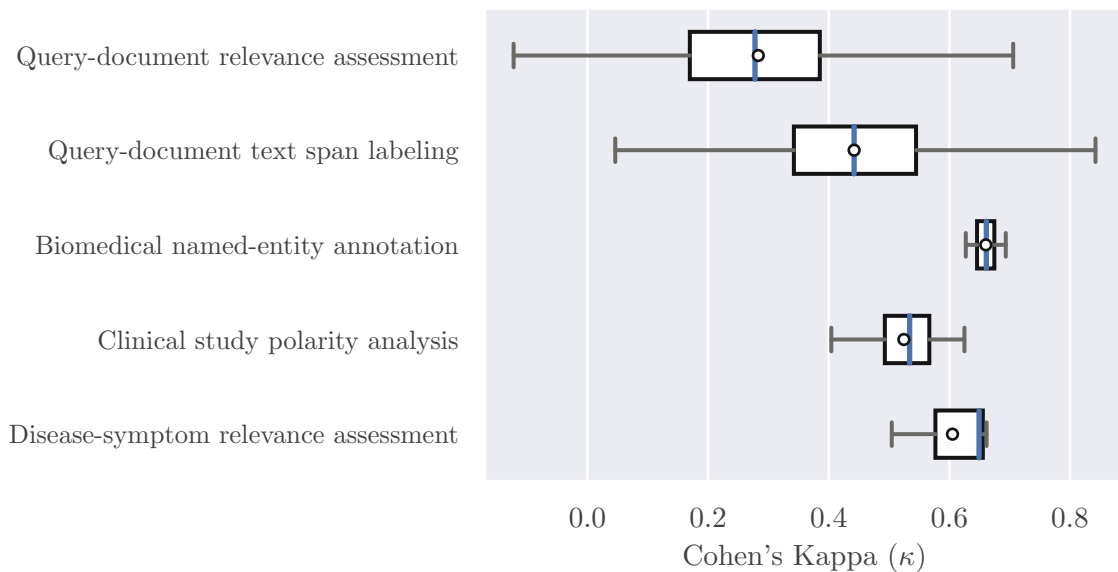


Figure 2.7: Cohen's Kappa agreements between pairs of annotators for the five tasks. The line in each box indicates the median agreement and the dot the average agreement.

### Subjectivity

A factor harming inter-annotator agreement is subjectivity. Strongly affected by subjectivity is the query-document relevance assessment task. For example, consider the query-document pair illustrated in Figure 2.8. The query asks about the Jamaican weather, and the document provides information about the weather in Jamaica. While one annotator might find the document relevant in answering the query, another annotator might find that the document misses relevant information such as degrees in Celsius or Fahrenheit. For the illustrated query-document pair, 44 annotators assigned the label *perfect* (i.e., fully answering the query) and 33 *partial* (i.e., partially answering the query). It is difficult to argue which group of annotators assigned the correct label since the core intent of the person who originally submitted the search query is unknown.

We aim to reduce the query-document relevance assessment task's subjectivity by simulating a 2-class setup instead of differentiating between 4-classes. Specifically, we combine the two classes *wrong* and *topic* as irrelevant and the two classes *partial* and *perfect* as relevant. After

**how is the weather in jamaica**

Jamaica Weather... When Is The Best Time To Visit? "Jamaica Weather... When Is The Best Time To Visit? A popular Jamaican poem starts by summing up Jamaica weather like this: We have neither Summer nor Winter Neither Autumn nor Spring We have instead the days When the gold sun shines on the lush green canefields- Magnificently And it's absolutely true. This is Jamaica weather! Most of our days are filled with warmth and sunshine, even during the rainy season. Jamaica has a tropical climate with hot and humid weather at sea level. The higher inland regions have a more temperate climate.

Figure 2.8: Sample of the task of query-document relevance assessment. This sample was labeled by 44 annotators as *perfect*, 33 as *partial*, 1 as *topic*, and 1 as *wrong*.

combining the classes, we re-compute the Cohen's Kappa agreement for the 2-class setup. We measure an average Kappa agreement of 0.59, which is a substantial improvement compared to the 0.28 agreement when differentiating between the four classes (see Figure 2.7). Although combining the classes improves the inter-annotator agreement and the dataset's reliability, we lose the fine-granularity of differentiating between four classes. Consequently, we would also limit evaluation and supervised training to the 2-class setup.

### Consistency

The next challenge analyzed is consistency in assigning text span labels. The labeling of text spans involves highlighting relevant words or characters within a given text. Depending on the length of a text, there are many different possibilities on how the text can be labeled. We collect text span annotations for two tasks: the labeling of PIC entities in clinical trial reports and the labeling of relevant text spans for query-document pairs. The average token length<sup>18</sup> is 293 tokens for clinical trial reports and 137 tokens for the documents. For each token, the annotator can assign a label, leading to a high number of possibilities on how a sample can be labeled. The many possibilities cause frequent disagreement between annotators.

We analyze the collected text span annotations by computing how frequently annotators label the exact same text spans for a sample (i.e., a document or a clinical trial report). For that, we pair annotators who labeled samples in common and computed how often they labeled exactly the same tokens or at least partially overlapping tokens. For the task of biomedical named-entity annotation, we find that annotators agree exactly in 11% of the cases and partially in 81%<sup>19</sup>. We observe similar results for labeling query-document pairs with an exact agreement of 9% and a partial agreement of 85%. For partial agreement, we found that annotators tend to disagree on multiple tokens, with an average of 26 disagreeing tokens for the relevance labeling task and 8 disagreeing tokens for labeling PIC. Although annotators rarely agree fully, the average Kappa agreement is moderate, with 0.66 for the PIC task and 0.44 for the document labeling task, as shown in Figure 2.7. The moderate agreement indicates a common conception among annotators on what tokens should be highlighted. However, the rare occurrence of the exact agreement shows that selecting token ranges in full agreement with other annotators poses a substantial challenge.

<sup>18</sup>The token length is computed based on the `word_tokenize` function of python's NLTK library.

<sup>19</sup>The remaining 8% are cases where the assigned token annotations do not overlap at all.

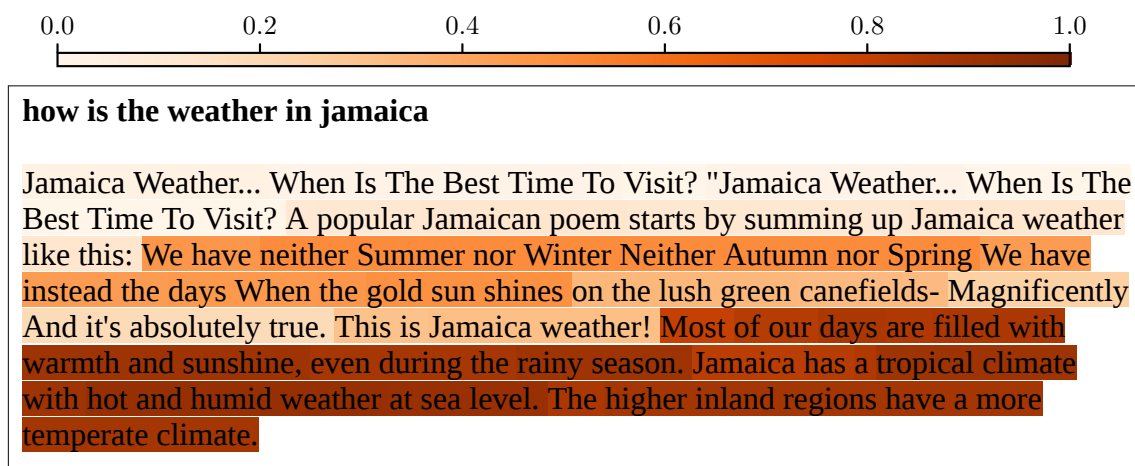


Figure 2.9: Sample of the task of query-document text span labeling with a heat-map indicating the tokens that were labeled by the 87 annotators

We conduct a case-based analysis of annotated token ranges for both text span annotation tasks. For the task of query-document text span labeling, we illustrate a document with labeled token ranges in Figure 2.9. Notice that most of the 87 annotators found the last two sentences of the document relevant, which describe Jamaica’s day-to-day weather for the query *how is the weather in jamaica*. Some annotators also labeled earlier text spans as relevant, which describe a poem about the weather of Jamaica. For the task of biomedical named-entity annotation, we illustrate the abstract of a clinical trial report with labels assigned for Participants in Figure 2.10. Notice that all five annotators labeled the phrase *46 adults undergoing elective inguinal hernia repair* in the middle of the abstract. The phrase *inguinal hernia surgery* at the beginning of the abstract was labeled by three annotators, and the phrase *elective inguinal hernia repair* appearing at the end by two. These two phrases mention a surgical procedure without explicitly mentioning the persons undergoing the procedure, which probably caused the disagreement. Note that to generate Figure 2.9 and Figure 2.10, we collected labels for a few samples from all annotators.

### Domain-Specificity

The last task-specific challenge analyzed is the difficulty of conducting annotation projects in the medical domain. Labeling data in the medical domain with a high agreement is challenging because of its jargon and terminology. A common approach to address this problem is to recruit expert annotators. We recruited expert annotators for the conducted medical-related tasks of polarity analysis, named-entity annotation, and disease-symptom relevance assessment. We reported average Cohen’s Kappa scores for these tasks of 0.52 for polarity analysis, 0.66 for named-entity annotation, and 0.61 for disease-symptom relevance assessment (see Figure 2.7). The inter-annotator agreement shows that even though we recruited medical experts, the annotators commonly disagree on how certain samples should be annotated.

We first analyze the disagreement of the biomedical polarity analysis task. For this task, annotators labeled each sentence appearing in a clinical trial report as either *positive*, *neutral*, or *negative*. We illustrate sentences labeled by the five expert annotators of this task in Table 2.6. The first three sentences are labeled in full agreement between the annotators as positive, neutral, and

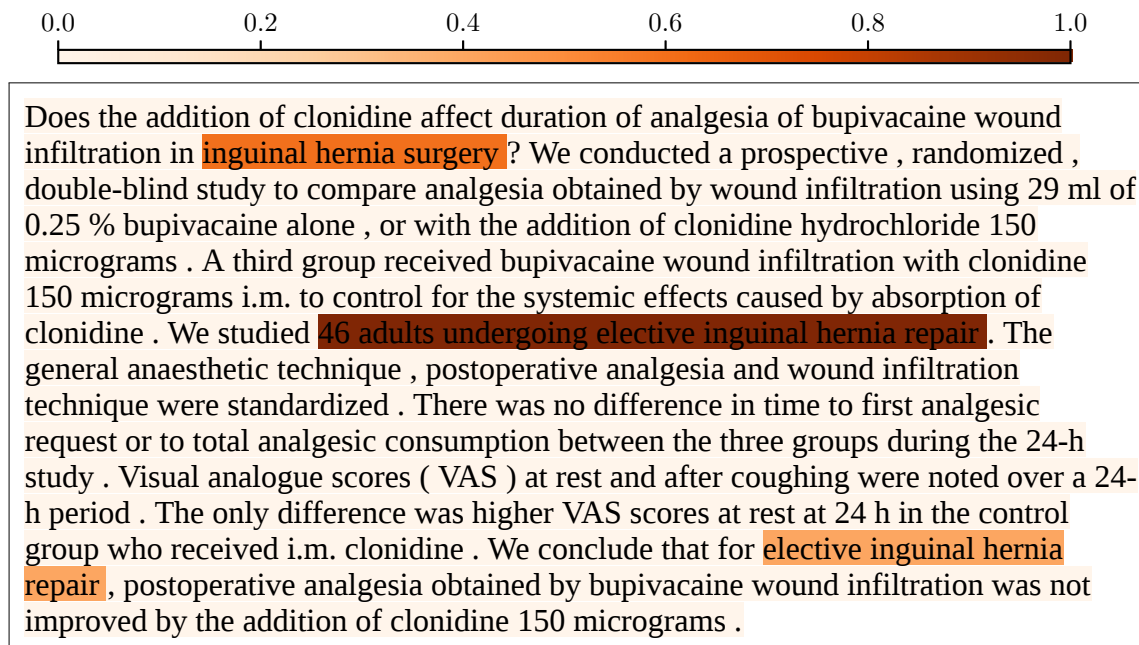


Figure 2.10: Sample of the task of biomedical named-entity annotation with a heat-map indicating the tokens that were labeled by the 5 annotators as Participants

negative, respectively. For the last two sentences, however, annotators did not have a common conception on what label should be selected, leading to disagreement. These last two sentences are difficult to label since they contain abbreviations and study-specific parameters. For example, consider the fourth sentence in Table 2.6, for which an annotator needs to know whether a lower *mean bond strength* is a positive or negative outcome of the study. Such jargon appears commonly in clinical trial reports and represents a core challenge of this task [NZLH05].

The PIC named-entity annotation task poses several challenges for annotators: First, the annotators need to understand the context of a clinical trial report to differentiate between the three labels accurately. For example, a drug can be part of a Population or part of an Intervention depending on the context, as shown in Table 2.7. Another challenge is the length of relevant phrases compared to traditional named-entity tasks such as labeling dates, movie names, or company brands. Especially, the phrases describing the Participants are often long and connected by several prepositions, e.g., *men with [...] from [...] after experiencing [...]*. The third and final challenge discussed is related to the medical jargon of clinical trial reports. For example, consider the phrase *oxycodone alone and combined with ethanol*. This phrase interlaces the Intervention (i.e., *oxycodone with ethanol*) and the Comparison (i.e., *oxycodone alone*), leading to uncertainty and disagreement on how this sample should be annotated. Although we aimed to cover such cases in our annotation guidelines, there are many other ambiguous cases similar to those presented. Addressing all special cases in the guidelines is impracticable, leading to lengthy instructions.

The third task for which we analyze challenges is the labeling of primary symptoms for disease-symptom pairs. We recruited three expert annotators for this task, who labeled 232 disease-symptom pairs for 20 diseases. We reported average Kappa agreements of 0.61 for this task, which shows that even the experts occasionally disagreed on labeling primary symptoms of the 20

## 2. BACKGROUND ON MANUAL CORPUS ANNOTATION

Table 2.6: Labeled sentences of the biomedical polarity analysis task. For each sentence, we report the number of labels assigned by the five expert annotators for Positive (Pos.), Neutral (Neu.), and Negative (Neg.).

| Sentence  | Pos. | Neu. | Neg. |
|---|------|------|------|
| Cognitive therapy had enduring effects that lasted beyond the end of treatment.   | 5    | 0    | 0    |
| Further investigation is needed to determine the effects on blood pressure and lipids.  | 0    | 5    | 0    |
| We conclude that for elective inguinal hernia repair , postoperative analgesia obtained by bupivacaine wound infiltration was not improved by the addition of clonidine 150 micrograms. | 0    | 0    | 5    |
| The results appear to indicate that mean bond strengths recorded in vivo following comprehensive orthodontic treatment are significantly lower than bond strengths recorded in vitro.   | 2    | 1    | 2    |
| Higher glucose levels after diagnosis were associated with a small but significantly higher BDI score and more ADM use.   | 2    | 2    | 1    |

Table 2.7: Depending on the context, treatment methods need to be labeled as Population or Intervention.

| Example   | Population                            | Intervention   |
|---|---------------------------------------|----------------|
| Adverse effects of aspirin in men who take vitamin C regularly  | men who take vitamin C regularly      | aspirin        |
| Adverse effects of vitamin C in men                             | men                                   | vitamin C      |
| Effects of paracetamol in patients who underwent bankart repair | patients who underwent bankart repair | paracetamol    |
| Bankart repair in patients with shoulder instability            | patients with shoulder instability    | Bankart repair |

diseases. We give an overview of the 20 diseases with Cohen’s Kappa scores for each in Table 2.8. Notice that the diseases are from different medical specialties: mental (e.g., *Depression*), dental (e.g., *Periodontitis*), digestive (e.g., *Appendicitis*), and respiration (e.g., *Asthma*). The variety in specialties makes labeling primary symptoms challenging, even for medical practitioners since they are usually not an expert in all the different specialties. The recruitment of annotators with omniscient expertise is infeasible when considering that there are about 135 medical specialties (Section 2.4.4), such as oncology, pathology, or cardiology. Although it would be possible to narrow the task to a certain medical specialty, this would limit the dataset’s application field.

Table 2.8: Average Cohen’s Kappa agreement with standard deviations per disease for the task of primary symptom labeling of disease-symptom pairs

| Disease               | Avg. $\kappa$<br>Agreement | Disease                            | Avg. $\kappa$<br>Agreement |
|-----------------------|----------------------------|------------------------------------|----------------------------|
| Mental Depression     | $0.15 \pm 0.37$            | Erysipelas                         | $0.69 \pm 0.10$            |
| Trigeminal Neuralgia  | $0.24 \pm 0.39$            | Epididymitis                       | $0.70 \pm 0.21$            |
| Migraine Disorders    | $0.36 \pm 0.18$            | Bronchitis                         | $0.74 \pm 0.18$            |
| Measles               | $0.39 \pm 0.13$            | Gastroesophageal reflux disease    | $0.74 \pm 0.18$            |
| Sleep Apnea Syndromes | $0.40 \pm 0.26$            | Asthma                             | $0.76 \pm 0.08$            |
| Myocardial Infarction | $0.46 \pm 0.15$            | Chronic Obstructive Airway Disease | $0.82 \pm 0.12$            |
| Periodontitis         | $0.48 \pm 0.17$            | Diabetes Mellitus                  | $0.84 \pm 0.12$            |
| Influenza             | $0.57 \pm 0.14$            | Pulmonary Embolism                 | $0.84 \pm 0.11$            |
| Cholecystitis         | $0.62 \pm 0.27$            | Anorexia Nervosa                   | $1.00 \pm 0.00$            |
| Tonsillitis           | $0.64 \pm 0.13$            | Appendicitis                       | $1.00 \pm 0.00$            |

## 2.6 Summary

This chapter presented the background on manual corpus annotations with respect to components, activities, and best practices. We divided the annotation effort, which we referred to as the annotation project, into the following four stages:

- **Project Definition:** This stage was about planning and defining an annotation project. We started by formulating a project goal as a concise and comprehensive statement of purpose, using not more than two sentences [PS12e]. Then, we described different types of annotating text, including document annotation, text span annotation, and linked text span annotation. Finally, we described the importance of studying related annotation projects to find and reuse best practices.
- **Data Preparation:** This stage was about acquiring and preparing a corpus. We described the importance of constituting the corpus from text data that is balanced and representative regarding the task at hand. Afterward, we described how a corpus is decomposed into smaller, self-contained units, defined as a *sample*. Finally, we discussed how annotations are best stored alongside the corpus and how large the corpus should be.
- **Execution:** This stage was about assigning the annotations to create the annotated corpus. As part of this stage, we described the annotators with respect to recruitment, motivation, compensation, and expertise. Moreover, we described how annotators are trained via the annotation guidelines and how they assign labels supported by an annotation tool. Finally, we described the annotation process, in which the annotators, the guidelines, and the annotation tool are involved to create the annotated corpus.
- **Evaluation:** For this stage, we described the inter-annotator agreement to measure the quality of annotations. We computed the inter-annotator agreement using the Cohen’s Kappa statistic, which is robust with respect to annotators agreeing by chance. Finally, we described common challenges that cause disagreement between annotators, including subjectivity, inconsistency, and domain-specificity.

We demonstrated the application of the four-stage annotation framework on five tasks. We make the created datasets publicly available for other researchers interested in using the data



## 2. BACKGROUND ON MANUAL CORPUS ANNOTATION

---

for evaluation or supervised learning. Furthermore, the datasets are a viable resource for other researchers studying cross-annotator behavior since each sample is labeled by multiple annotators. The annotated datasets for the task of relevance assessment and text span labeling for query-document pairs can be found on GitHub<sup>20</sup>, with the annotation interface being available at<sup>21</sup>. The datasets for PIC annotation and polarity analysis are also available on GitHub<sup>22</sup>, including a list of 44 person keywords used to pre-label semantic entities in the clinical trial reports (described in Section 2.5.1). The dataset for disease-symptom annotation is available for non-commercial purposes after signing a non-disclosure agreement<sup>23</sup>.

We use the presented four-stage annotation framework as the foundation for the annotation projects conducted in the upcoming chapters of this thesis. Each project is described with respect to its task, goal, corpus, annotators, annotation guidelines, and annotation tool. In Chapter 5, we collect annotations with the aim of improving the time efficiency for labeling large volumes of data. In Chapter 6, we collect annotations with the aim of improving the accuracy of non-expert annotators for domain-specific tasks.

---

<sup>20</sup><https://github.com/sebastian-hofstaetter/fira-trec-19-dataset/>

<sup>21</sup><https://github.com/pkerschbaum/fira>

<sup>22</sup><https://github.com/Markus-Zlabinger/kconnect>

<sup>23</sup>Contact the author of this thesis if you are interested in accessing the dataset



# State-of-the-Art

This thesis aims to improve the efficiency and effectiveness of manual data acquisition processes through unsupervised text similarity methods. In this chapter, we present related work on the three core topics of the thesis: unsupervised text similarity, efficient text annotation, and effective text annotation. We begin by reviewing related work on improving the efficiency of the manual data acquisition process with respect to time and costs (Section 3.1). Then, we describe literature on improving the effectiveness of the data acquisition process by increasing the label accuracy of human annotators (Section 3.2). Finally, we review related work on unsupervised text similarity in Section 3.3.

## 3.1 Efficient Text Annotation

This section reviews related work on improving the time- and cost-efficiency of the human data annotation process. Although plenty of research improves the annotation efficiency for computer vision tasks, such as labeling images or videos [MBS09, YNC08], only limited literature exists for the IR and NLP domain. Transforming ideas and approaches for image or video annotation to text data is often not possible since the computer-vision approaches aim to assist the workers for problems specific to visual data, such as the annotation of object boundaries in images. Therefore, we limit the review of related work on efficient text annotation approaches.

A common approach to obtain text annotations time-efficiently is to use crowdsourcing platforms [SBDS14]. By publishing annotation tasks on crowdsourcing platforms, an enormous base of workers has access to the task and can create annotations. In other words, the annotation task when using crowdsourcing is carried out by many individual workers in a collaborative effort. While this approach is clearly time-efficient, the cost-efficiency depends on the payment per annotated sample, which is usually set by the annotation practitioner. Setting the payment too low leads to an unfair payment for workers [STL<sup>+</sup>18], who might decide not to participate in the task at all. On the other hand, an excessive payment leads to surplus project costs, making the annotation procedure cost-inefficient. Overall, the availability of a large worker base and lower payment compared to employing experts allows annotation practitioners to obtain massive amounts of annotations time- and cost-efficiently through crowdsourcing [She18, SBDS14].

Snow et al. [SOJN08] evaluated the efficiency of crowd-annotations for five NLP tasks. The five tasks being: affect recognition, word similarity judgment, word sense disambiguation, event temporal ordering, and recognizing textual entailment. They collected crowd-annotations from the Mechanical Turk platform for each task and computed the inter-annotator agreement to a set of expert annotations. The crowd-annotations reached high agreements to experts for all tasks, especially when multiple redundant annotations are aggregated into a higher quality meta-annotation. Snow et al. conclude that crowd-annotations are a suitable alternative for the five tasks, with the advantage of being collected more time- and cost-efficiently than expert annotations.

Another research direction to reduce the time and cost used for data annotation is active learning. In the process of active learning, only a sub-set of automatically selected samples is labeled, namely those samples that are highly informative. For computing the informativeness of samples, several selection criteria are proposed in the literature [FZL13]. The most commonly used criterion is uncertainty sampling, where a continuously trained supervised learning model selects a sample that lies on the prediction boundary to be annotated next. The main research direction regarding active learning is the development of new selection criteria [FZL13, ZWTM10]. Another direction is the combination of active learning with the aforementioned crowdsourced annotation [LSS11, FYT14]. The core benefit of using active learning is that the data annotation process is time- and cost-efficient since only a sub-set consisting of highly informative samples is manually labeled.

Neubig et al. [NM10] propose the efficient annotation approach of *Partial Annotation*. The idea behind partial annotation is to automatically identify important parts within a text sample and then label only the important parts instead of the entire sample. They evaluated their approach on the task of Japanese pronunciation estimation and developed a point-wise estimator to identify ambiguous words important for annotation. By labeling only the important words rather than the entire sentences, they showed a substantial improvement in time efficiency while preserving a high label quality of the obtained annotations.

Seifert et al. [SUKG13] describe an efficient approach for document annotation. They propose to condense full-text documents to only the key sentences and key phrases appearing in the document via an unsupervised approach. Since the key information represents the critical parts of a document, only this condensed information is manually labeled. Their evaluation compares the annotation of full-text documents to the annotation of only the condensed key information. They show that texts containing only the condensed information can be annotated twice as fast while preserving a similar coverage for labeling relevant information compared to the annotators who labeled the full-text documents.

Another strategy to improve the efficiency of manual annotation is using a semi-automatic approach. Semi-automatic approaches consist of two steps: First, annotations are assigned by an automatic system, which are then manually verified by humans. Manually verifying annotations is usually conducted more quickly than assigning annotations from scratch, improving the manual labeling procedure's time efficiency. The strategy of combining automatic annotation with manual verification has been successfully applied for various tasks, such as entity linking [DDC12a], part of speech tagging [MMS93], and biomedical information retrieval [NIDL11].

The presented related work on efficient text annotation can be roughly categorized as (i) crowdsourcing-based approaches, (ii) approaches for selectively labeling samples or sample parts that are highly informative, and (iii) combining automatic annotation with manual verification. This thesis proposes a novel direction for improving the time- and cost-efficiency of text annotation tasks. In our approach, similar samples are annotated in groups to reduce the effort for annotators

to cognitively process the samples. This approach of a GROUP-WISE annotation is described in Chapter 5, including experiments for the task of question-answering and named-entity recognition.

## 3.2 Effective Text Annotation

Here, we review related work on effective text annotation regarding the label quality of annotators. Reaching a certain quality threshold is challenging, especially for tasks that require specific expertise to be performed. An important step in preparing annotators for a difficult task is the training of annotators [DKC<sup>+</sup>18]. The training usually includes providing task instructions as a comprehensive description of how the task should be performed [NLP<sup>+</sup>18, SOJN08]. In addition to task instructions, also a few examples are often provided that demonstrate in a practical way how the task is performed [JSPW17]. Providing task instructions and demonstration examples is critical for an effective annotation task design [DKC<sup>+</sup>18], especially when the annotators are not familiar with the task, as is often the case when recruiting annotators from crowdsourcing platforms.

Several studies have researched the concept of providing demonstration examples to annotators. Doroudi et al. [DKBH16] study the effectiveness of training non-expert annotators for difficult tasks. One evaluated approach was to show examples labeled by experts to non-expert annotators recruited from a crowdsourcing platform. They show that training crowdworkers based on the examples annotated by experts is highly effective compared to various other training strategies. Singla et al. [SBB<sup>+</sup>14] provide examples specific to the currently annotated sample: For an image labeling task, a machine learning approach was used to dynamically select relevant examples from an expert-authored set based on the progression of each worker. Liu et al. [LSB<sup>+</sup>16] propose an annotation task design called *Gated Instructions* to improve the quality of annotators. The Gated Instructions consist of several strategies to improve worker accuracy, including an interactive tutorial, worker feedback, and regular screening for low-accuracy workers. They apply the Gated Instructions approach to the task of relation extraction and show a substantial increase in worker accuracy. The increased accuracy of individual workers resulted in an annotated dataset of high quality, demonstrated by training and evaluating machine learning algorithms on the dataset. The reviewed literature on providing examples to annotators shows the importance of examples as an essential part of an effective text annotation.

Various strategies for effective data annotation are summarized by Daniel et al. [DKC<sup>+</sup>18]. They describe strategies such as combining the labels from several workers for the same sample into a higher-quality annotation by aggregation (e.g., majority voting). Furthermore, they describe critical aspects impacting the label quality, such as:

- the annotation interface (e.g., should be intuitive);
- the recruitment of workers (e.g., test run to find accurate workers);
- the task instructions (i.e., should be clear, understandable, and unambiguous);
- and the regular screening for low-accuracy workers (e.g., by mixing test samples into the annotation procedure).

As another critical factor, Daniel et al. [DKC<sup>+</sup>18] highlight the intrinsic and extrinsic motivation of workers to perform the task with high accuracy. The core motivational incentives for annotators can be divided into three categories: personal, social, and financial [PCK<sup>+</sup>13]. The primary motivation in annotation projects, especially when utilizing crowdsourcing, is a fair payment with

the possibility of rewarding annotators that do exceptionally well on a task [STL<sup>+</sup>18]. To foster intrinsic motivation, practitioners of annotation tasks can design the task to be fun, educating, or with a greater purpose (e.g., collecting a dataset for skin cancer prediction) [RKK<sup>+</sup>11].

A technique specific for increasing the intrinsic motivation of annotators is Games with a Purpose (GWAP) [JN14, HB12, VAD04], where the annotation task is designed as a game. Poesio et al. [PCK<sup>+</sup>13] propose *Phrase Detectives*, a game for anaphora annotation. The game contains classical design concepts such as levels, scores, and leaderboards. Their study aimed not only to improve the label quality of individual annotators but also to give the workers the chance to validate existing annotations. The authors report that over 2.5 million judgments were collected from almost 8,000 players. Madge et al. [MCKP17] propose *TileAttack*, a game with the purpose of collecting annotations for the NLP task of text segmentation. *TileAttack* is a game where players are awarded points based on marked tokens. The game aims to increase accuracy and player engagement for the task of text segmentation. The reviewed related work reports a positive impact on the effectiveness of data annotation tasks when using gamification approaches. However, designing annotation tasks as a game is complex, and not every task is suitable to be transformed into a fun and engaging game [PCK<sup>+</sup>13].

Another factor that impacts the effectiveness of an annotation task design is the complexity of the task. Finnerty et al. [FKTC13] show that a task's cognitive complexity affects both accuracy and time efficiency. Cheng et al. [CTIB15] increase the output quality and worker experience by splitting large tasks into smaller tasks. They provide evaluation results for the following three tasks: sorting, transcription, and arithmetic calculations. Feyisetan et al. [FSL<sup>+</sup>18] conduct annotation experiments for named-entity recognition in tweets and found that the length and number of entities in a tweet influenced the quality of the crowd-annotations: A better quality was obtained for shorter tweets with fewer entity mentions. Sabou et al. [SBDS14] recommend as a best practice in corpus annotation to keep the text samples that are annotated reasonably short, without compromising the context.

We reviewed related work on an effective text annotation for various topics, including games with a purpose, effective annotation task design, and training annotators via task instructions and demonstration examples. This thesis proposes a new approach for presenting demonstration examples to annotators. Specifically, we propose the DEXA approach, where demonstration examples are dynamically retrieved from a set of expert samples based on their similarity to the currently annotated sample. The related work closest to our approach is the image labeling task of Singla et al. [SBB<sup>+</sup>14], where similar images are shown as examples to support annotators. Our approach is based on a similar principle but adopts an unsupervised text similarity method to find relevant expert examples instead of creating an internal machine learner model. As another core difference, we aim to support non-expert workers for text annotation tasks that usually require specific expertise to be conducted correctly. We describe the DEXA approach in Chapter 6, where we show the effectiveness of the approach based on a complex named-entity annotation task of the biomedical domain.

### 3.3 Unsupervised Text Similarity Methods

The annotation approaches proposed in this thesis use an unsupervised text similarity method. Computing the similarity between two texts is a fundamental topic of the research fields IR and NLP [BR99]. For these two fields, we review the development of new methodologies starting from traditional word-based methods to more recent contextualized text embedding methods. Then,

we review related work on the evaluation of unsupervised similarity methods, where we describe benchmark corpora commonly used by the research community to report novel advancements.

### 3.3.1 Unsupervised Text Similarity Methods

Two well-known unsupervised similarity methods, often used for generating baseline results in the IR field, are the TFIDF and BM25 weighting schema [BR99]. The TFIDF weighting schema computes the similarity between texts based on the term frequency (TF) and inverse document frequency (IDF), defined as the term's overall frequency across an entire corpus. The Okapi Best Matching 25 (BM25) weighting schema incorporates, in addition to term-frequency and inverse document-frequency, also the document length since longer documents have a higher chance to contain terms in common with a search query. The weighting function of the BM25 method is more effective than the TFIDF method and therefore reaches higher evaluation scores on various benchmark corpora [MRS08]. Studies reporting new scientific advancements in IR often report baseline results using the BM25 and the TFIDF weighting schema. These methods are well-known by the community, which allows an intuitive interpretation when comparing the baseline results with the results of new methodologies.

The research directions of IR and NLP substantially changed with the introduction of word embeddings by Mikolov et al. [MSC<sup>+</sup>13]. The underlying idea behind word embedding is that a word is not represented by its character representation but by a distributed vector based on the context in which a word usually appears. Consider the following phrase with a missing word in the middle:

*the furry cat \_\_\_\_\_ over the box*

Knowing the missing word's context, we can predict that the word is probably something like *jumps*, *crawls*, or *falls*. Based on this principle of predicting words from their context, we can generate distributed word vectors from large text corpora using supervised machine learning. These distributed word vectors are known as word embeddings and are used to compute the similarity between words. Words that appear in a similar context (such as *jumps* and *crawls*) will have a higher similarity than words that appear in different contexts. We summarize the idea behind word embeddings by quoting the famous linguist John Rupert Firth, who said [JR57]:

*"You shall know a word by the company it keeps"*

For generating word embeddings, two algorithms are commonly used. First, the *word2vec* algorithm proposed by Mikolov et al. [MSC<sup>+</sup>13, LM14], and second, the *Glove* algorithm proposed by the Stanford NLP group [PSM14]. Both approaches generate word embeddings from words appearing in large text corpora, which can be problematic when inferring word embeddings for words that did not appear during the generation procedure. To address this problem of so-called out-of-vocabulary words, Bojanowski et al. [BGJM17] proposed the *fastText* algorithm, which also uses character n-grams to generate word embeddings. Through word embeddings, new methodologies for IR and NLP emerged, including new approaches for the task of unsupervised text similarity, as described next.

Based on word embeddings, new methods were proposed to compute the unsupervised similarity between texts. These methods usually consist of the following three steps [LM14, ALM17]: First, an input text is tokenized. Second, the word embedding for each token in the text is inferred. Finally, the individual word embeddings are aggregated into a vector as a representation of the

input text. The similarity between two aggregated vectors can be computed using vector distance functions such as the cosine similarity [ALM17]. The research focus on unsupervised similarity based on word embeddings is on the aggregation step. Simple aggregation methods such as averaging were proposed [LM14]. More recent methods perform a weighted aggregation based on the importance of individual words [ALM17].

The recent trend in unsupervised text similarity methods has shifted towards contextualized text embeddings [PGJ18, CKS<sup>+</sup>17]. These methods directly infer a vector representation for an input text by incorporating the order of the terms appearing in the text. The incorporation of word order is an improvement to the aforementioned method of aggregating word embeddings, where the word order is not considered. In the literature, two approaches are usually used to compute the similarity between two contextualized text embeddings [PGJ18]: First, the similarity is predicted by a supervised machine learning algorithm, and second, the similarity is computed via the cosine similarity of the two text embeddings. Note that the prediction of the similarity via supervised machine learning requires the availability of sufficient data for model training. Methods for computing contextualized text embeddings are usually based on complex neural network architectures containing millions of parameters [CYK<sup>+</sup>18, CKS<sup>+</sup>17]. The neural networks are trained in an unsupervised way based on large text corpora using network architectures such as the Bidirectional Encoder Representations from Transformers (BERT) [DCLT18]. Using contextualized word embeddings has led to state-of-the-art performances for various NLP-related tasks, including sequence tagging, sentence classification, dependency parsing, question-answering, or natural language inference [BLC19, DCLT18, LYK<sup>+</sup>19].

### 3.3.2 Evaluation of Unsupervised Similarity Methods

We evaluate unsupervised similarity methods ranging from traditional methods to recent text embedding methods as one contribution of this thesis. This section will review related work on the evaluation of unsupervised similarity methods, focusing on benchmark corpora and comparative studies. We limit the review of related work to the two tasks relevant to this thesis: question-to-question similarity and biomedical sentence similarity.

#### Question-to-Question Similarity

One part of this thesis is the retrieval and grouping of similar questions so that these question groups can be annotated efficiently. The retrieval of similar questions is also an essential topic in the research area of Community Question Answering (CQA). In CQA online forums, such as Stackoverflow or Quora, users ask questions, which are then answered by the community. A common problem of CQA forums is duplicated questions, which are redundant questions that were already asked and answered before. The retrieval and automatic identification of duplicated questions are critical research problems in the area of CQA [HEABH17, CLG16, FKS16].

A duplicate question retrieval task was part of the Semantic Evaluation (SemEval) workshop [NMM<sup>+</sup>16, NHM<sup>+</sup>17]. In the scope of the SemEval workshop series, new benchmark datasets are released to advance the state-of-the-art of various NLP tasks. Research teams participating in the workshop compete to reach the best performance results on the benchmark datasets. In the scope of the SemEval workshop held in 2016 [NMM<sup>+</sup>16] and 2017 [NHM<sup>+</sup>17], a competitive task for similar question retrieval was conducted. The task for the participating teams was to retrieve relevant questions for a set of candidate questions. The questions were sampled from the Qatar Living CQA forum, where users can exchange information about various topics of Qatar, including lifestyle, news, and events.



Table 3.1: State-of-the-art results reported for the SemEval 2016 [NMM<sup>+</sup>16] and 2017 [NHM<sup>+</sup>17] similar question retrieval dataset.

|                         | Mean Average Precision (MAP) |              |
|-------------------------|------------------------------|--------------|
|                         | SemEval 2016                 | SemEval 2017 |
| Charlet et al. [CD17]   | 0.80                         | 0.48         |
| Hazem et al. [HEABH17]  | 0.79                         | 0.49         |
| Goyal [Goy17]           | -                            | 0.47         |
| Filice et al. [FDSMM17] | 0.79                         | 0.49         |

The similar question benchmark datasets of SemEval 2016 and 2017 were publicly released after the workshop and are now a standard benchmark for evaluating question-to-question similarity methods. The evaluation metric commonly reported is Mean Average Precision (MAP), measuring the quality of ranking-based systems. We give an overview of state-of-the-art results reported for the 2016 and 2017 SemEval benchmark dataset in Table 3.1. The reported methodologies are based on supervised learning using training data released by the SemEval workshop organizers. Evaluation results for unsupervised methods [ZW18] are reported scarcely, often only for traditional methods as a baseline result (e.g., TFIDF as a baseline in [Goy17]). The literature lacks a comparative and comprehensive evaluation of unsupervised similarity methods for similar question retrieval.

### Biomedical Sentence Similarity

The second reviewed task is about computing the semantic similarity of texts in the biomedical domain. Computing the similarity between texts of the biomedical domain is essential to quickly search and find relevant information within the massive number of medical papers available [Her09]. Alone in the open-access archive PubMed Central (PMC)<sup>1</sup>, there are 6.5 million articles available as of October 2020. Many benchmark datasets have been released for the task of biomedical information retrieval [VH12] and biomedical question-answering [PRLP18]. However, for this thesis, we are interested in biomedical benchmark datasets specific to computing the similarity between sentences.

There are two standard benchmark datasets commonly used for evaluating biomedical sentence similarity methods. The first dataset, published by Wang et al. [WAF<sup>+</sup>18], is the Medical Semantic Text Similarity dataset (MedSTS). The second dataset, published by Soğancıoğlu et al. [SÖÖ17], is the BIOSSES dataset. Both datasets contain sentence pairs for the task of computing the similarity between the sentence pairs. The sentence pairs are manually annotated by medical experts with an integer value from 0 (not similar) to 4 (highly similar) for BIOSSES and from 0 (not similar) to 5 (highly similar) for MedSTS.

The MedSTS and BIOSSES dataset are frequently used to measure the performance of biomedical sentence similarity methods. The standard evaluation metric is the Pearson correlation coefficient, measuring the correlation between the manual annotations and the similarity scores computed by an automatic system. We report the Pearson correlation coefficient of state-of-the-art results for both datasets in Table 3.2. The reported results are based on contextualized text embedding methods (e.g., BERT and ELMo), combined with techniques such as transfer learning, multitask

<sup>1</sup><https://www.ncbi.nlm.nih.gov/pmc/>

learning, and systematic fine-tuning. The listed approaches are all supervised by using a subset of the data to train a machine learning model to predict the similarity scores. Although results for individual unsupervised methods are occasionally reported (e.g. [CPL19, TS20]), no comprehensive evaluation of unsupervised methods is available for the two benchmark datasets.

### Comparative Studies for Evaluating Unsupervised Similarity Methods

For this thesis, we require a comparative evaluation of unsupervised text similarity methods for the two tasks of biomedical sentence similarity and question-to-question similarity. Standard benchmark datasets are available for both tasks, including the aforementioned SemEval CQA dataset [NHM<sup>+</sup>17] and the two biomedical sentence-to-sentence similarity datasets BIOSSES [SÖÖ17] and MedSTS [WAF<sup>+</sup>18]. For all three benchmark datasets, results are usually reported for *supervised* methods using the training data accompanying each dataset [WAL<sup>+</sup>18b, NHM<sup>+</sup>17, CPL19]. On the other hand, only a few papers report results for *unsupervised* similarity methods. Papers reporting unsupervised results are usually from one of two categories: (i) papers that consider traditional unsupervised methods such as TFIDF to generate baseline results (e.g., [RG19, Goy17]) and (ii) papers that propose a new unsupervised method and report results for this specific method (e.g., [ZCY<sup>+</sup>19, CPL19]). Only Tawfik et al. [TS20] compared several unsupervised methods on the BIOSSES and MedSTS benchmark dataset. However, their study is limited to contextualized embedding methods without considering traditional methods or methods based on aggregating word embeddings.

We fill the research gap of a comparative study of unsupervised similarity methods by evaluating ten methods in Chapter 4 for the task of sentence similarity and question-to-question similarity. We evaluate traditional methods (e.g., TFIDF [BR99]), methods based on aggregated word embeddings (e.g., averaged aggregation [LM14]), and methods based on contextualized text embeddings (Sentence-BERT [RG19]). As benchmark datasets, we consider the BIOSSES dataset, the MedSTS dataset, the SemEval CQA dataset, and a new dataset from the customer-support domain for question-answering. Our comparative evaluation analyzes the effectiveness of the different methods, allowing researchers to quickly identify effective methods for their task at hand.

Table 3.2: State-of-the-art results reported for the BIOSSES [SÖÖ17] and MedSTS [WAL<sup>+</sup>18b] sentence similarity dataset.

|                                 | Pearson Correlation Coefficient |         |
|---------------------------------|---------------------------------|---------|
|                                 | MedSTS                          | BIOSSES |
| Peng et al. [PYL19]             | 0.85                            | 0.92    |
| Chen et al. [CPL19]             | 0.84                            | 0.85    |
| Gu et al. [GTC <sup>+</sup> 20] | -                               | 0.92    |
| IBMResearch*                    | 0.90                            | -       |
| CBI*                            | 0.90                            | -       |
| UFL*                            | 0.89                            | -       |

\* Top-3 performing teams on the n2c2/OHNLP shared task 2019 [WFS<sup>+</sup>20]: IBMResearch (IBM Corporation), NCBI (National Center for Biotechnology Information), UFL (University of Florida).



## 3.4 Summary

This chapter reviewed related work on the three core topics of this thesis: efficient text annotation, effective text annotation, and unsupervised text similarity. We discussed how unsupervised text similarity methods evolved from traditional word-based methods to contextualized text embedding methods. We then described related work on evaluating unsupervised similarity methods, focusing on the task of question-to-question similarity and sentence similarity in the biomedical domain, as these are the same tasks as in our data annotation experiments. We highlight the research gap of a comparative evaluation of unsupervised similarity methods on publicly available benchmark datasets for both domains. We address this gap in Chapter 4, where we conduct a thorough evaluation of ten unsupervised semantic similarity methods on four datasets of the biomedical and question-answering domain.

The second topic reviewed in this section was the efficiency of text annotation. For this topic, we reviewed related work that aims to improve the time- and cost-efficiency of the annotation process. We reviewed work from the areas of crowdsourcing and active learning. Afterward, we discussed a few papers that automatically identify the essential parts in a text so that only these parts require human annotation. In Chapter 5 of this thesis, we propose a novel direction for a time- and cost-efficient annotation by annotators labeling groups of similar samples instead of labeling samples one by one.

The third topic reviewed was the effectiveness of text annotation. For that, we reviewed related work that improves the accuracy and label quality obtained from annotators. We described various techniques for an effective annotation, including, e.g., the training of annotators, the creation of intuitive annotation interfaces, and the preselection of accurate workers through test runs. We further described approaches from the area Games with a Purpose, where the annotation task is designed as a game to foster the intrinsic motivation of annotators. In Chapter 6, we propose a new approach for effective annotation, where we show examples to crowdworkers that are similar to the currently annotated sample and therefore provide dynamic support to annotators in their decisions.



Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.  
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

# Evaluation of Unsupervised Short-Text Similarity Methods

In this thesis, we use unsupervised semantic short-text similarity (SSTS) methods to support human workers at data annotation tasks. During our search for an effective, unsupervised SSTS method, we observed a lack of comparative evaluations in the literature. A comparative evaluation, however, is the prerequisite for an informed decision on an effective method based on empirical evidence.

In this chapter, we perform a comparative evaluation of ten SSTS methods, ranging from traditional count-based methods (e.g., TFIDF) to recent contextualized embedding methods (e.g., Sentence-BERT [RG19]). We evaluate the effectiveness of three different preprocessing functions and 13 publicly available language models for word and text embeddings. The evaluated language models were pre-trained on domain-specific corpora (such as biomedical publications) and general-purpose corpora (such as Wikipedia, web news, online forums).

We evaluate the ten SSTS methods based on four benchmark corpora. The corpora are diverse, consisting of

- the languages English and German
- the tasks sentence-to-sentence similarity and similar question retrieval
- the domains medical, technical customer support, and Community Question Answering

The characteristics of the benchmark corpora are similar to the data that we will use in our annotation experiments (Chapter 5 and 6), and therefore, the obtained results for the ten SSTS methods allow us to select the most effective methods for our use cases. Beyond our use cases, the versatility of the methods and the benchmark datasets make the obtained results useful to other researchers intending to use unsupervised SSTS methods in fields such as question-answering or ad-hoc information retrieval.

The contributions of this chapter are as follows:

- We perform a comparative evaluation of ten unsupervised SSTS methods. We report results for a broad range of publicly available pre-trained models and test the impact of different preprocessing functions.
- We compare the effectiveness of the methods based on four benchmark corpora coming from different tasks, languages, and domains.

The remainder of this chapter is structured as follows: The unsupervised SSTS methods are summarized in Section 4.1. We describe the pre-trained models and the preprocessing functions in Section 4.2. The experiments and results for the task of medical sentence-to-sentence similarity are reported in Section 4.3. The experiments and results for the task of similar question retrieval are described in Section 4.4. We summarize this chapter in Section 4.5.

## 4.1 Unsupervised Short-Text Similarity Methods

We first describe the ten unsupervised SSTS methods that are evaluated. Each method computes a similarity score

$$\text{sim}(t, k) \in \mathbb{R} \quad (4.1)$$

between two short-texts  $t, k$ . The similarity score is an indicator of the semantic similarity between  $t$  and  $k$ , where a high score indicates a similar meaning between both texts, and a low score suggests a dissimilar meaning.

We compute the similarity score for methods that derive a vector representation  $\mathbf{v} \in \mathbb{R}^n$  using the cosine similarity. The cosine similarity between two vectors  $\mathbf{v}_t$  and  $\mathbf{v}_k$  representing the texts  $t$  and  $v$  is defined as

$$\text{sim}(t, k) = \cos(\mathbf{v}_t, \mathbf{v}_k) = \frac{\mathbf{v}_t \cdot \mathbf{v}_k}{\|\mathbf{v}_t\| \cdot \|\mathbf{v}_k\|} \in \mathbb{R}, \quad (4.2)$$

where  $\|\mathbf{v}\|$  is the Euclidean norm of a vector  $\mathbf{v}$ . The cosine similarity ranges from  $-1$  (dissimilar) to  $+1$  (similar) and is a standard measure to compute the similarity between vectorized texts.

The ten evaluated methods can be categorized as follows: methods that are based on (i) word counts, (ii) aggregated word embeddings, and (iii) text embeddings. We describe each category and the associated methods next.

### 4.1.1 Word Count Based Methods

These methods compute the similarity between two texts based on their words in common. Methods from this category are successors of the traditional Bag-of-Words model [BR99] (BOW), where a text is represented as a multiset (i.e., a *bag*) of words. As a result of representing texts as a multiset of words, two semantically similar words with a different character representation, such as "like" and "enjoy", are considered dissimilar. We evaluate the following two methods from this category:

**TFIDF** [BR99]: This method derives weighted word vectors  $\mathbf{v}_t, \mathbf{v}_k$  for two texts  $t$  and  $k$ . The word vectors are computed based on the term frequency (TF) and the inverse document frequency (IDF) of the words appearing in  $t$  and  $k$ . The IDF is a weighting schema that indicates the importance of a word according to the word's overall occurrence within a given text corpus: Words that rarely appear in the corpus are weighted as more important than words that appear frequently.

**Levenshtein Distance** [MRS08]: The similarity between two texts is computed based on the number of edits to change the word sequence of  $t$  into the word sequence of  $k$ . The edits are deletion, insertion, and substitution of words. The number of edits to transform  $t$  into  $k$  is defined as the distance  $d(t, k)$ . Distance is a measure of dissimilarity rather than similarity, and therefore, we transform  $d(t, k)$  into a similarity measure  $\text{sim}(t, k)$  as follows: First, we compute the maximum word length of  $t$  and  $k$ , defined as  $\max(|t|, |k|)$ , where  $|t|$  and  $|k|$  is the number of words in  $t$  respectively  $k$ . Note that the maximum word length is also the upper bound of edits to transform  $t$  into  $k$ , meaning that  $\max(|t|, |k|) \geq d(t, k)$ . Afterward, we normalize the distance by the maximum word length and subtract it from 1 to obtain the similarity score  $\text{sim}(t, k) = 1 - \frac{d(t, k)}{\max(|t|, |k|)}$ , ranging from 0 (no similarity) to 1 (high similarity). The Levenshtein method is the only method of the evaluated ones that computes a distance rather than a similarity score.

#### 4.1.2 Aggregated Word Embedding Methods

A word embedding is the representation of a word as a dense, high-dimensional vector [MSC<sup>+</sup>13]. Word embeddings are mathematically defined as an  $n$ -dimensional vector representation  $\mathbf{e}_w \in \mathbb{R}^n$  of a word  $w$ . Based on the vector representations, we can compute the semantic similarity between words using the cosine similarity. The computed similarity depends on the context in which a word appears: If two words usually appear in a similar context, the cosine similarity between the embedded word vectors is high. The advantage of using word embeddings over the traditional BOW model is that we can measure the semantic similarity between words even though they have a different character representation, such as "like" and "enjoy".

Word embeddings cannot be used out-of-the-box to compute the semantic similarity between texts since a prior aggregation is necessary. During aggregation, the individual word embeddings of a text are used to generate a text embedding  $\mathbf{v} \in \mathbb{R}^n$ , for example, by averaging the individual word vectors. After aggregation, the semantic similarity between two text embeddings  $\mathbf{v}_t$  and  $\mathbf{v}_k$  is computed using the cosine similarity from Equation 4.2. We compare three methods for aggregating word embeddings:

**Average Embedding (AVG)** [LM14]: A basic approach to aggregate a set of word embeddings into a text embedding is by averaging. Averaging the word embeddings of a text  $t$  is defined as  $\text{avg}(t) = \frac{1}{|t|} \sum_{w \in t} \mathbf{e}_w$ , where  $\mathbf{e}_w$  is the embedding vector for word  $w$  and  $|t|$  is the number of words in  $t$ . The similarity between two texts  $t$  and  $k$  is computed as  $\cos(\text{avg}(t), \text{avg}(k))$ , where  $\cos$  is the cosine similarity, defined in Equation 4.2.

**Weighted Average Embedding (WAVG)** [LM14]: In the AVG method, each embedding vector is weighted equally, neglecting the degree of importance of words. To incorporate the word importance, we can weigh each word vector by its TFIDF value. The TFIDF weighted aggregation for a text  $t$  is defined as  $\text{wavg}(t) = \frac{1}{|t|} \sum_{w \in t} \mathbf{e}_w \times \text{tfidf}(w)$ , where  $\text{tfidf}(w)$  is the TFIDF weight for a word  $w$ . The similarity between two texts  $t$  and  $k$  is computed as  $\cos(\text{wavg}(t), \text{wavg}(k))$ .

**Smooth Inverse Frequency (SIF)** [ALM17]: This method aggregates the embeddings of a text  $t$  following a two-step approach: First, a weighted average embedding vector is computed as  $\text{sif}(t) = \frac{1}{|t|} \sum_{w \in t} \frac{a}{a+p(w)} \mathbf{e}_w$ , where  $a$  is a hyper-parameter<sup>1</sup>, and  $p(w) = \frac{tf(w)}{|W|}$  is the relative term frequency of word  $w \in W$  across all texts  $t$  in a corpus  $T$ . In the second step, the weighted average vectors for all texts  $(t_1, \dots, t_y) \in T$  are organized as row vectors in a matrix  $M = (\text{sif}(t_1), \dots, \text{sif}(t_y))$ . From  $M$ , the projection of the first principal component is removed, resulting in the matrix  $M_{pca}^-$ . The principal component removal can be considered as a form of denoising [ALM17]. The similarity between two texts  $t_i, t_j \in T$  is the cosine similarity between the row vectors  $i$  and  $j$  of  $M_{pca}^-$ .

### 4.1.3 Text Embedding Methods

Methods from this category directly infer a text embedding  $\mathbf{v}_t \in \mathbb{R}^n$  for an input text  $t$ . In contrast to aggregated word embeddings and word count-based methods, this category’s methods compute a contextualized word embedding by encoding the word position when inferring a vector representation. By computing contextualized embeddings, two texts  $t$  and  $k$  containing the same words but in a different order will yield two different vector representations  $\mathbf{v}_t$  and  $\mathbf{v}_k$ . Consequently, two texts containing the same set of words in a different order can result in a cosine similarity  $\leq 1$ . For methods from the category word count based and aggregated word embedding, the similarity of two texts with the same words is always 1, regardless of word order. Encoding word order can be crucial to capture a text’s meaning accurately. For example, consider the two sentences "A *low* number of patients survived after *high* intensity laser therapy." versus "A *high* number of patients survived after *low* intensity laser therapy".

We consider text embedding methods that are compatible with sentences and questions, as these are the short-text types used in our experiments (see Section 4.3 and 4.4). The following four text embedding methods are evaluated:

**Doc2Vec** [LM14]: This method is an extension to the word embedding algorithm Word2Vec towards embedding documents such as phrases, questions, sentences, and paragraphs. While Word2Vec [MSC<sup>+</sup>13] computes vector representations of words, Doc2Vec directly infers a dense vector  $\mathbf{v}_t$  representing the input text  $t$ . When inferring a text embedding, only those words in  $t$  are considered that also appear in the training corpus. Words that do not appear in the training corpus, so-called out-of-vocabulary words, are ignored.

**Sent2vec** [PGJ18]: A common strategy to derive text and word embeddings is to use the Continuous Bag of Words model (CBOW), in which words are predicted based on their immediate context (i.e., the surrounding words). Sent2vec extends the CBOW model by first splitting words into character n-grams, followed by predicting the n-grams based on their immediate context. The incorporation of character n-grams makes this method robust to out-of-vocabulary words during the inference of a vector representation  $\mathbf{v}_t$ .

**Sentence-BERT (SenBERT)** [RG19]: The Bidirectional Encoder Representations from Transformers (BERT) is a language model leading to state-of-the-art results for many NLP tasks [DCLT18]. However, text representations derived from the BERT model are ineffective in computing an unsupervised text similarity via the cosine similarity [RG19]. This problem is addressed in SenBERT, which uses a BERT-based siamese network to create independent sentence representations  $\mathbf{v}_t$  suitable for unsupervised tasks.

<sup>1</sup>We set the hyper-parameter  $a$  to  $10^{-3}$ , as suggested in [ALM17].

**Universal Sentence Encoder (USE)** [CYK<sup>+</sup>18]: The USE model employs simple Transformers to encode sentences as dense vector representations  $\mathbf{v}_t$ . The model is pre-trained on a variety of text corpora, such as texts from Wikipedia, web news, and online forums. This versatility of model training makes the inferred text embeddings  $\mathbf{v}_t$  universally applicable for various tasks, including clustering, text classification, and semantic short-text similarity.

**InferSent** [CKS<sup>+</sup>17]: InferSent creates sentence embeddings  $\mathbf{v}_t$  based on a Bidirectional Long Short-Term Memory (BiLSTM) neural network trained on natural language inference (NLI) data. The NLI data used to train the network is the Stanford NLI corpus, introduced in [BAPM15]. Similar to the USE method, InferSent embeddings are designed to be universally applicable to various tasks, including semantic short-text similarity.

## 4.2 Preprocessing and Pre-trained Language Models

We evaluate the SSTS methods with respect to three preprocessing functions: *Identity* where no preprocessing is conducted, *Lower* where text is lowercased, and *LowerStop* where text is lowercased and stopwords are removed. We use the English stopword list of the NLTK Python library<sup>2</sup>. For the tokenization needed for the methods Levenshtein, TFIDF, AVG, WAVG, and SIF, we use the *word\_tokenize* function of the NLTK library. Note that we do not report exhaustive preprocessing results for all methods since certain methods expect (i) a specific preprocessing to be effective (e.g., lowercasing for TFIDF), or (ii) the raw unprocessed text as input, as it is the case for SenBERT, USE, and InferSent.

Methods based on word and text embeddings require a language model trained on large amounts of text data. The text data used in the training procedure is optimally the same text data on which the similarity methods are evaluated. However, for some benchmark corpora, large amounts of text data for training are not available, making corpus-specific training from scratch infeasible. Furthermore, the model training from scratch is cumbersome since a fine-tuning of hyper-parameters is necessary, and a high amount of computational resources is required for training the language model.

The alternative to corpus-specific model training is to use publicly available pre-trained models. These models are often trained on public domain corpora such as Wikipedia articles [BGJM17] or open-access scientific publications [BLC19]. The advantage of using pre-trained models is that they are applicable out-of-the-box to infer word or text embeddings. Another advantage in using pre-trained models is reproducibility since the models can be downloaded and re-used by other researchers to reproduce the results of an experiment.

The pre-trained models used in our experiments are summarized in Table 4.1. All described models are freely available, and more details on the models can be found in the referenced papers, including download links, hyper-parameter settings, and descriptions of the text corpora used for training. We select models that are either pre-trained on (i) biomedical data (e.g., PubMed, MIMIC III), (ii) general English corpora (e.g., Wikipedia, web news), or (iii) multilingual data (e.g., translation text pairs). We preferably select these models since they are pre-trained on data similar to the data of the four benchmark corpora on which we evaluate the SSTS methods. For example, one considered benchmark corpus is in German, for which we use the multilingual models. As another example, two benchmark corpora are based on text data from biomedical publications, for which we use the models pre-trained on text data from biomedical literature.

<sup>2</sup><https://www.nltk.org/> (version 3.5)

Table 4.1: Overview of the evaluated pre-trained models

| Category            | Model  | Training Data   | Used by Method |
|---------------------|--|---|----------------|
| Word<br>Embedding   | BioWord2Vec [ZCY <sup>+</sup> 19]                        | PubMed abstracts, MIMIC III corpus [JPS <sup>+</sup> 16]          | AVG, WAVG, SIF |
|                     | PubMedW2VSmall <sup>a</sup> [CCKP16]                     | PubMed abstracts  | AVG, WAVG, SIF |
|                     | PubMedW2VLarge <sup>b</sup> [CCKP16]                     | PubMed abstracts  | AVG, WAVG, SIF |
| Text<br>Embedding   | WikiDoc2Vec [LB16]                                       | English Wikipedia   | Doc2Vec        |
|                     | SciBERT [BLC19]  | Semanticscholar full-text papers                                  | SenBERT        |
|                     | DistBERT-Multi [RG20]                                    | SNLI corpus [BAPM15], STS benchmark corpus [CDA <sup>+</sup> 17]  | SenBERT        |
|                     | BioBERT [LYK <sup>+</sup> 19]                            | PubMed abstracts  | SenBERT        |
|                     | ClinicalBERT [AMB <sup>+</sup> 19]                       | MIMIC III corpus [JPS <sup>+</sup> 16]                            | SenBERT        |
|                     | USE 4.0 [CYK <sup>+</sup> 18]                            | Wikipedia, web news, online forums, SNLI corpus [BAPM15]          | USE            |
|                     | USE-Multi 3.0 [YCA <sup>+</sup> 19]                      | question-answering pairs, translation pairs, SNLI corpus [BAPM15] | USE            |
|                     | InferSent 2.0 [CKS <sup>+</sup> 17]                      | SNLI corpus [BAPM15]  | InferSent      |
| BioSent2Vec [CPL19] | PubMed abstracts, MIMIC III corpus [JPS <sup>+</sup> 16] | Sent2Vec  |                |

<sup>a</sup> Training window size 2

<sup>b</sup> Training window size 30

### 4.3 Medical Sentence Similarity

We evaluate the ten unsupervised similarity methods for the task of medical sentence-to-sentence similarity. For this task, sentence pairs are given, and for each pair, a ground truth label is available, indicating the semantic similarity between the two sentences. The ground truth labels are based on the manual judgment of medical experts. The aim of this task is to compute a similarity score automatically—in our case, by using the ten SSTs methods—between each sentence pair to best approximate the human judgment.

#### 4.3.1 Experiment Setup

We consider two sentence-to-sentence similarity benchmark corpora for our experiments:

- **BIOSSES** [SÖÖ17]: This corpus contains 100 sentence pairs with labeled similarity scores from 0 (not similar) to 4 (highly similar). The sentences are sampled from biomedical research papers.
- **MedSTS** [WAF<sup>+</sup>18]: This corpus contains 1,068 sentence pairs annotated from 0 (not similar) to 5 (highly similar). The sentences are sampled from anonymized electronic health records of patients of the Mayo Clinic.



We compute the effectiveness of the ten unsupervised SSTS methods via the Pearson correlation coefficient between the manually assigned ground truth labels and the score computed by the unsupervised methods. The Pearson correlation is the standard metric reported for these two corpora and is defined as

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4.3)$$

where:  $n$  is the number of sentence pairs,  $x_i$  is the human judgment for the sentence pair at index  $i$ ,  $y_i$  is the score computed by the unsupervised similarity methods, and  $\bar{x}$ ,  $\bar{y}$  are the arithmetic means.

We compute the Pearson correlation coefficient over all samples of each corpus instead of splitting the datasets into a training and testing set. Training data is unnecessary since the evaluated methods are unsupervised without requiring any corpus-specific training. Furthermore, we do not perform any corpus-specific training of language models since such training is infeasible given the small corpus size of BIOSSES and MedSTS.

### 4.3.2 Evaluation of Effectiveness

The evaluation results in Table 4.2 show the high effectiveness of methods that use the BioSent2Vec or BioWord2Vec model. These models' common denominator is their pre-training on biomedical research papers (i.e., PubMed) and clinical notes (i.e., the MIMIC III corpus), which is similar to the underlying data source of BIOSSES and MedSTS.

Apart from the pretraining, the method also has a substantial impact on the obtained results. The SenBERT method, although pre-trained on biomedical publications, is rather ineffective, even outperformed by TFIDF-weighted word vectors. Similarly ineffective are the universal methods USE and InferSent. These findings align with other studies that report that transformer-based text representations are highly effective as input for supervised learning but less effective in an unsupervised setting [RG19, TS20].

The effect of preprocessing shows that stopword removal is usually beneficial, especially for Levenshtein and AVG, since these two methods do not have an incorporated mechanism for weighting word importance.

## 4.4 Similar Question Retrieval

In this section, we describe our experiments on evaluating the ten SSTS methods for the task of similar question retrieval. This task is defined as follows: Given a query question, the aim is to find relevant questions that have the same intent as the query question and rank them as highly as possible. This task is empirically evaluated based on ground truth labels indicating the relevance between query questions and the other questions.

#### 4. EVALUATION OF UNSUPERVISED SHORT-TEXT SIMILARITY METHODS

Table 4.2: Pearson correlation coefficient for the task of biomedical sentence-to-sentence similarity. The Pearson correlation is computed between the ground truth labels and the similarity score of the unsupervised SSTS methods. For each corpus, we highlight the overall best result **bold** and the best result per category by underline.

| Category       | Method         | Model          | Preprocessing | MedSTS      | BIOSSES     | Avg.        |
|----------------|----------------|----------------|---------------|-------------|-------------|-------------|
| Word count     | TFIDF          | -              | Lower         | <u>0.74</u> | 0.70        | 0.72        |
|                | TFIDF          | -              | LowerStop     | <u>0.74</u> | <u>0.73</u> | <u>0.74</u> |
|                | Levenshtein    | -              | Lower         | 0.55        | 0.64        | 0.60        |
|                | Levenshtein    | -              | LowerStop     | 0.64        | 0.69        | 0.66        |
| Word embedding | AVG            | BioWord2Vec    | Lower         | 0.61        | 0.72        | 0.66        |
|                | AVG            | BioWord2Vec    | LowerStop     | 0.72        | <b>0.77</b> | 0.75        |
|                | AVG            | PubMedW2VSmall | Lower         | 0.45        | 0.65        | 0.55        |
|                | AVG            | PubMedW2VSmall | LowerStop     | 0.64        | 0.76        | 0.70        |
|                | AVG            | PubMedW2VLarge | Lower         | 0.48        | 0.63        | 0.55        |
|                | AVG            | PubMedW2VLarge | LowerStop     | 0.66        | 0.76        | 0.71        |
|                | WAVG           | BioWord2Vec    | Lower         | 0.73        | 0.75        | 0.74        |
|                | WAVG           | BioWord2Vec    | LowerStop     | 0.76        | <b>0.77</b> | 0.76        |
|                | WAVG           | PubMedW2VSmall | Lower         | 0.64        | 0.74        | 0.69        |
|                | WAVG           | PubMedW2VSmall | LowerStop     | 0.68        | 0.76        | 0.72        |
|                | WAVG           | PubMedW2VLarge | Lower         | 0.67        | 0.74        | 0.70        |
|                | WAVG           | PubMedW2VLarge | LowerStop     | 0.71        | <b>0.77</b> | 0.74        |
|                | SIF            | BioWord2Vec    | Lower         | <u>0.79</u> | 0.75        | <u>0.77</u> |
|                | SIF            | BioWord2Vec    | LowerStop     | 0.78        | 0.76        | <u>0.77</u> |
|                | SIF            | PubMedW2VSmall | Lower         | 0.71        | 0.75        | 0.73        |
|                | SIF            | PubMedW2VSmall | LowerStop     | 0.70        | 0.76        | 0.73        |
| SIF            | PubMedW2VLarge | Lower          | 0.72          | 0.75        | 0.74        |             |
| SIF            | PubMedW2VLarge | LowerStop      | 0.71          | 0.76        | 0.73        |             |
| Text embedding | Doc2Vec        | WikiDoc2Vec    | Lower         | <b>0.81</b> | 0.75        | 0.78        |
|                | Doc2Vec        | WikiDoc2Vec    | LowerStop     | 0.80        | 0.76        | 0.78        |
|                | Sent2Vec       | BioSent2Vec    | Lower         | <b>0.81</b> | 0.74        | 0.78        |
|                | Sent2Vec       | BioSent2Vec    | LowerStop     | <b>0.81</b> | <b>0.77</b> | <b>0.79</b> |
|                | SenBERT        | SciBERT        | Identity      | 0.60        | 0.68        | 0.64        |
|                | SenBERT        | BioBERT        | Identity      | 0.78        | 0.58        | 0.68        |
|                | SenBERT        | ClinicalBERT   | Identity      | 0.65        | 0.69        | 0.67        |
|                | USE            | USE 4.0        | Identity      | 0.66        | 0.72        | 0.69        |
| InferSent      | InferSent 2.0  | Identity       | 0.49          | 0.65        | 0.57        |             |

### 4.4.1 Experiment Setup

We consider two similar question retrieval benchmark corpora for our experiments:

- **Customer-Support Questions:** This corpus consists of 500 German customer-support questions originating from an automatic Q-A system of a telecommunication company. For each of the 500 questions, ground truth labels are available, indicating the relevance to the remaining 499 questions. Two questions are considered relevant if they have the same intent, or in other words, the same need for information. The questions were randomly sampled from a set of 113,394 questions, asked between February and July 2016 by customers of the telecommunication company. The questions were asked to the company’s chat-bot system, which tries to map questions to an answer set. The questions are usually of two types: First, questions concerning technical problems, such as connection disturbances, forgotten credentials, or a locked phone. And second, questions seeking general information on topics like the contract, webmail, or roaming fees. The average question length is 3.18 words, and the median length is 2 words (both computed when ignoring the stop words).
- **SemEval 2017:** The second considered corpus is the benchmark dataset of the SemEval 2017 question-to-question similarity task (task 3 subtask B) [NHM<sup>+</sup>17]. The dataset consists of candidate questions and 10 related questions per candidate, manually labeled as either relevant or irrelevant with respect to the candidate. The goal of the task is to re-rank the related questions. We use the test set of the SemEval 2017 dataset for our experiments, which contains 88 candidate questions. The questions in the dataset stem from the Qatar Living Community Question Answering forum and are in English. The average length of the questions is 3.09 words, and the median is 3 words.

We evaluate the ten SSTS methods based on the two corpora by computing the Mean Average Precision (MAP). We consider MAP since it is a standard metric for evaluating information retrieval systems producing a ranked list of results. The MAP is defined as

$$\text{MAP} = \frac{\sum_{q=1}^Q \text{AveP}(q)}{Q}, \quad (4.4)$$

where  $Q$  is the number of queries and  $\text{AveP}(q)$  is the average precision of the query  $q$ .

We train corpus-specific models from scratch for both benchmark corpora. We do so since the two corpora contain user-generated questions affected by colloquial language or spelling mistakes – often not considered in pre-trained models from well-formatted text corpora such as Wikipedia. For the model training, we have large amounts of training data available for both the Customer-Support Question corpus (113,394 additional questions) and the SemEval 2017 corpus, where we use the training dataset provided by the organizers of the SemEval task. We create the following models for each set of training data: Doc2Vec, Sent2Vec, and a fastText model used for aggregated word embeddings. Note that we used fastText rather than the Word2Vec model since it is robust with respect to out-of-vocabulary words. The robustness to out-of-vocabulary words is crucial considering the characteristics of user-generated questions (e.g., misspellings) and questions in the German language (e.g., compound words<sup>3</sup>).

<sup>3</sup>Words in German can be compounds from multiple words, which increases the uniqueness of words and makes out-of-vocabulary words more likely.

Table 4.3: Mean Average Precision (MAP) for the task of similar question retrieval. For each corpus, we highlight the overall best result **bold** and the best result per category by underline.

| Category       | Method      | Model                               | Preprocessing | Customer Support | SemEval 2017 | Avg.        |
|----------------|-------------|-------------------------------------|---------------|------------------|--------------|-------------|
| Word count     | TFIDF       | -                                   | Lower         | <u>0.30</u>      | <u>0.40</u>  | <u>0.35</u> |
|                | TFIDF       | -                                   | LowerStop     | <u>0.30</u>      | 0.39         | <u>0.35</u> |
|                | Levenshtein | -                                   | Lower         | 0.15             | 0.32         | 0.24        |
|                | Levenshtein | -                                   | LowerStop     | 0.20             | 0.32         | 0.26        |
| Word embedding | AVG         | -                                   | LowerStop     | 0.40             | 0.40         | 0.40        |
|                | WAVG        | Corpus Trained                      | LowerStop     | 0.43             | 0.43         | 0.43        |
|                | SIF         | -                                   | LowerStop     | <b>0.46</b>      | <b>0.44</b>  | <b>0.45</b> |
| Text embedding | Doc2Vec     | Corpus Trained                      | LowerStop     | 0.41             | <u>0.43</u>  | 0.42        |
|                | Sent2Vec    | Corpus Trained                      | LowerStop     | <b>0.46</b>      | 0.42         | <u>0.44</u> |
|                | SenBERT     | DistBERT-Multi [RG20]               | Identity      | 0.31             | 0.41         | 0.36        |
|                | USE         | USE-Multi 3.0 [YCA <sup>+</sup> 19] | Identity      | 0.37             | 0.42         | 0.40        |

We use the following setup for model training: The raw text is preprocessed using the *LowerStop* preprocessing function. For the fastText model, we select the default hyper-parameters as described in [BGJM17]. For the Doc2Vec model, we set the vector size to 300 and the window size to 3, as these parameters are well suited for most tasks [LM14]. For all other parameters not explicitly mentioned, we use the defaults of the Gensim [RS10] (fastText, Doc2Vec), GitHub (Sent2Vec<sup>4</sup>), and Scikit-learn [PVG<sup>+</sup>11] (TFIDF) implementations. We refer to the models trained from scratch as *corpus trained*.

#### 4.4.2 Evaluation of Effectiveness

Based on the Customer-Support Question dataset and the SemEval 2017 dataset, we compare the semantic similarity methods by Mean Average Precision in Table 4.3. The results show that the SIF method combined with a corpus trained language model is the most effective method for unsupervised similar question retrieval on both datasets. In contrast, the best reported supervised system of the SemEval 2017 workshop achieved only a slightly higher MAP of 0.472 [NHM<sup>+</sup>17], thus showing the high effectiveness of unsupervised methods for this task.

The pre-trained multilingual models for SenBERT and USE perform rather poorly in comparison to the corpus trained models. This finding aligns with our previous results on biomedical sentence-to-sentence similarity (Section 4.3) and other studies that evaluated transformer-based text representations in an unsupervised setting [RG19, TS20].

The use of different preprocessing functions for TFIDF and Levenshtein shows only slight improvements in MAP. We observed that many questions do not contain stopwords in the first place since users prefer to formalize questions as concisely as possible by expressing only the core intent. For example, users usually asked "webmail login page" when they needed information on the webmail service's login URL.

<sup>4</sup><https://github.com/epfml/sent2vec>

## 4.5 Summary

We performed a comparative evaluation of ten unsupervised SSTS methods. The methods were from three categories: methods based on word counts, aggregated word embeddings, and contextualized text embeddings. We compared the methods based on four benchmark corpora, coming from different domains, languages, and tasks. The tasks were sentence-to-sentence similarity and similar question retrieval. We evaluated various pre-trained models and preprocessing steps. The reported evaluation results regarding the various SSTS methods, pre-trained models, and preprocessing steps provide a decision basis for other researchers who seek an effective method for their use cases.

Our results showed that the Sent2Vec method and methods based on aggregated word embeddings are effective on all four benchmark corpora. When using aggregated word embeddings, a weighted method, such as WAVG or SIF, should be considered since these are superior to a simple average aggregation, as in the AVG method. Furthermore, weighted aggregated word embeddings and Sent2Vec text embeddings do not require a specific preprocessing function to be effective: We found that these methods perform equally effectively when the raw text is input, the text is lowercased, or the text is lowercased with stopwords removed. Therefore, we suggest using these methods with raw text as input without any special preprocessing.

We further found that the TFIDF method of the word count based category is a suitable choice for quick prototyping. Although this method does not lead to state-of-the-art results, it can be quickly implemented since it does not rely on a complex language model, as required by methods from the category aggregated word embeddings and text embeddings. We recommend using the TFIDF method with lowercasing and stopword removal as preprocessing steps, which slightly improves the method's effectiveness.

Finally, we describe how the evaluated methods are used in the remainder of this thesis. In the experiments where we compute the unsupervised similarity between biomedical sentences (Chapter 5 and 6), we select the Sent2Vec method with the pre-trained BioSent2Vec model. We select this combination of method and model since it was the most effective one on the biomedical sentence-to-sentence benchmark corpora. In the experiments where we retrieve similar questions to be annotated as groups (Chapter 5), we select the SIF method with corpus trained fastText embeddings. We select this setting since it was the most effective one for similar question retrieval and worked well on German texts, which is the same language in which we will annotate questions in groups.

We make the implementation of this chapter's experiments publicly available on GitHub<sup>5</sup>. To access the benchmark corpora BIOSSES and SemEval 2017, the download information is available in the corresponding papers [NHM<sup>+</sup>17] and [SÖÖ17], respectively. Access to the MedSTS corpus is granted for scientific purposes by contacting the first author of [WAF<sup>+</sup>18]. The corpus of customer-support questions was provided to us in the scope of a collaboration with a telecommunication company. Unfortunately, we are not permitted to share this corpus.

<sup>5</sup><https://github.com/Markus-Zlabinger/ssts>



Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.  
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

# Efficient Group-Wise Data Annotation

Corpora containing a high volume of annotated data represent a fundamental resource for training supervised learning algorithms. A high volume of annotated data is even more critical for training the data-hungry deep learning algorithms that emerged as an indispensable part of IR and NLP in the past decade. Despite the existence of publicly available corpora, such as the Stanford Question Answering Dataset (SQuAD) [RZLL16] or the Microsoft MACHine Reading COmprehension Dataset (MS MARCO) [BCC<sup>+</sup>16], a lack still exists for domain-specific tasks and languages other than English.

In cases where no appropriate annotated corpus is available, a new one can be created through manual annotation. However, manually annotating a corpus is usually a time-consuming and, therefore, costly procedure. The common approach for corpus labeling is to have annotators go through each data sample (e.g., sentences, questions, phrases) one by one and assign the correct label. We refer to this approach as *Sequential Annotation* (SEQUENTIAL).

This chapter proposes the GROUP-WISE annotation approach to create new corpora for IR and NLP-related tasks time-efficiently. In the GROUP-WISE approach, semantically similar samples are pre-grouped, allowing annotators to process these similar samples more quickly. The similar samples are grouped using an unsupervised semantic similarity method. Annotating a group of similar samples is especially time-efficient for tasks where label selection is laborious. For example, consider the task of assigning answer labels to questions, and the catalog in which the answer labels are looked up contains hundreds or even thousands of entries. Looking up the answer label that fits the currently labeled question is usually a time-consuming procedure. However, by grouping similar questions that have the same information need, the annotator might be able to re-use a looked-up answer to annotate multiple questions. We evaluate for the described task of question-answering the time efficiency of the GROUP-WISE approach.

The second task on which we evaluate the GROUP-WISE approach is annotating named-entities in the sentences of biomedical publications. For this task, we group sentences based on their semantic similarity into bundles of three and present these bundles to the human workers for labeling. The grouping of similar sentences into bundles positively affects efficiency: First, the

overhead for workers to switch between the different named-entity labels is reduced since an already selected label can often be re-used for annotating entities in several sentences. Second, the effort to cognitively process the three already similar samples is reduced, allowing annotators to assign labels more quickly.

We systematically compare the GROUP-WISE to the SEQUENTIAL annotation approach for the tasks question-answering and named-entity annotation. Our results show for the question-answering task that annotators using the GROUP-WISE approach assign labels 41% faster and require 51% fewer answer lookups than annotators of the SEQUENTIAL approach. For the named-entity annotation task, the annotators of the GROUP-WISE approach are 29% faster and require 16% fewer interactions than annotators of the SEQUENTIAL approach. Furthermore, we analyze the approaches from the perspective of label quality by computing Kappa agreements to a set of gold standard annotations. We find that the label quality is not affected (positively nor negatively) when using the GROUP-WISE annotation approach.

The contributions of this chapter are the following:

- We propose the GROUP-WISE annotation approach to create new annotated corpora for IR and NLP-related tasks efficiently. We thoroughly examine the effects of the proposed approach from the perspectives of time efficiency and label quality.
- We compare the GROUP-WISE annotation approach to the commonly used SEQUENTIAL approach based on the tasks question-answering and named-entity recognition.

The remainder of this chapter is structured as follows: We formally define and describe the GROUP-WISE annotation approach in Section 5.1. The application of the GROUP-WISE approach to the question-answering task is described in Section 5.2. The application to the named-entity recognition task is described in Section 5.3. We summarize the chapter in Section 5.4.

## 5.1 Group-Wise Annotation Approach

In the GROUP-WISE annotations approach, semantically similar samples are pre-grouped before being labeled by the human workers. The intuition behind this approach is that semantically similar samples often require a similar annotation, which allows workers to process the samples more time-efficiently. More formally, we define the GROUP-WISE approach given a set of unlabeled text samples  $S$  (e.g., questions, sentences, phrases) as follows:

1. A candidate sample  $s \in S$  is randomly selected.
2. All samples in  $S$  (except  $s$ ) are ranked with respect to  $s$ , using an unsupervised semantic similarity method.
3. The candidate and the most similar samples are grouped. We use two different grouping strategies, which will be described in more detail later.
4. The grouped samples are annotated.
5. The annotated samples are removed from the sample pool  $S$ , and a new iteration is started at *Step 1*, until all samples are labeled, namely when  $S = \{\}$ .

We require a strategy to group similar samples for *Step 3*. We use two strategies: *Automatic* and *Manual* grouping. The strategies are suited for different use-cases, as described next.



## Automatic Grouping

This strategy automatically generates groups by bundling each candidate with the corresponding top- $N$  most similar samples retrieved by the unsupervised semantic similarity method. Each sample in a generated group is labeled by the annotator individually. Although samples are labeled individually, the labeling is more time-efficient since, in many cases, the grouped samples are semantically similar, allowing annotators to process them quickly. Furthermore, the individual labeling makes this strategy compatible with text span annotation tasks (Section 2.1.2) since these tasks require labels to be associated with precise text parts of each sample. The main advantage of the automatic grouping strategy is that the time-effort to annotate one group can be estimated since each group contains an equal number of samples. Therefore, the automatic strategy is an optimal choice when the annotators are paid a fixed amount per labeled sample, which is usually the case when acquiring annotations from crowdsourcing platforms such as Mechanical Turk. We use the automatic grouping strategy for our experiments of annotating named-entities in biomedical publications.

## Manual Grouping

For this strategy, the annotator skims through the most similar samples and manually selects samples to be grouped with the candidate. The annotator groups only those samples that require the exact same labeling as the candidate, allowing her/him to assign one or multiple labels to the entire group. Assigning labels to the entire group makes this strategy the optimal choice for document annotation tasks (Section 2.1.2), such as labeling questions with answer labels. Another characteristic of the manual grouping strategy is that the time-effort for labeling a group of similar samples is difficult to estimate due to the manual selection involved: The manual selection is time-intensive if many samples are similar to the candidate and less time-intensive if only a few samples are similar. Therefore, the manual grouping strategy is suitable when the annotators are paid a fixed hourly wage, which is usually the case when collecting in-house annotations. We use the manual grouping strategy for the question-answering annotation task, described next.

## 5.2 Question-Answer Annotation

Personal information services via phone, e-mail, or live chat are a major cost factor for customer-oriented companies. To reduce these costs, companies implement question-answering (Q-A) systems so that users can obtain information autonomously. The evaluation and the supervised training of a deployed system depends on the availability of ground truth data.

A typical property of many Q-A systems is that questions are asked redundantly by different users. As a result, many questions refer to the same information need provided by a corresponding answer. The described scenario is typical in various Q-A domains such as tourism [PM16], telecommunication [Wan10], and medical [WYC05], as well as in search engines [WNZ02]. To create a ground truth dataset in such cases, annotation practitioners commonly use the SEQUENTIAL approach by annotating the questions one by one and label each with the relevant answer. Clearly, this approach is highly time-intensive when considering that for each question, the annotator has to lookup the answer in an answer catalog, which could potentially contain hundreds of entries.

We address this efficiency problem using the GROUP-WISE annotation approach to label questions with the same intent in groups. Questions have the same intent if they seek the same answer. The GROUP-WISE annotation approach for question-answering following the procedure of Section 5.1 consists of the following steps: First, a candidate question is selected, and the most similar

questions to it are presented to the annotator. Second, from the presented questions, the annotator selects the questions with the same intent as the candidate question. Finally, the annotator labels the entire group, consisting of the candidate question and the selected same-intent questions, with an answer. By grouping similar questions, looked-up answers can be assigned to the entire group of questions. As a result, fewer answer lookups are required, making the annotation procedure more time-efficient.

### 5.2.1 Experiment Setup

This section describes our experiment setup to evaluate the GROUP-WISE approach for question-answer annotation. We describe the dataset and the annotation tool used by the human workers.

#### Dataset & Task

The data used for our experiments consists of 500 customer-support questions from an Austrian telecommunication company<sup>1</sup>. The annotation task is to label the questions with an answer label. The answer label is selected from the Frequently Asked Question (FAQ) catalog of the company, containing 373 entries. Each answer covers a specific question commonly asked by the company's customers; therefore, multi-label assignments are not required for this annotation task. The questions and the answers are in German.

The questions originate from a chat-bot system, available through the company's website. Customers can consult this system to obtain answers to their questions autonomously. After asking a question, the system tries to map it to the correct answer. Acquiring manual annotation is crucial for evaluating the deployed chat-bot system and improving it through supervised machine learning.

#### Annotation Tool

The customer-support questions are labeled via an annotation tool that we specifically developed for this experiment. The tool incorporates the GROUP-WISE approach and is illustrated in Figure 5.1. The tool consists of the following core components:

- (A) The candidate question is presented to the annotator.
- (B) The ranked list of questions semantically similar to the candidate is shown. The annotator skims through this list and marks questions with the same intent as the candidate.
- (C) This component allows the annotator to perform answer lookups within the answer catalog. We provide a search functionality based on string matching to find answer labels quickly.
- (D) The annotator inputs the label ID of the answer to the current question. After pressing the *Annotate* button, the candidate and all marked questions (green background) are annotated with the input label ID. In cases where no answer ID is available in the catalog, the ID "-1" can be input to indicate a *no-answer* label for the current question.

The retrieval of similar questions is performed using an unsupervised semantic short-text similarity (SSTS) method. In our experiments, we select the Smooth Inverse Frequency (SIF) method to compute the similarity between questions. We select this method since it is highly effective for

---

<sup>1</sup>Note that we used the same dataset to evaluate unsupervised similarity methods in Chapter 4.

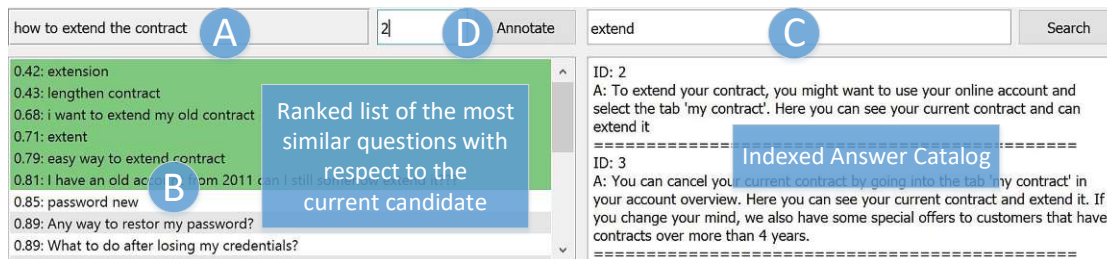


Figure 5.1: Annotation interface of the GROUP-WISE annotation tool for question-answering. The illustration shows: (A) the candidate question, (B) the ranked list of similar questions, (C) the answer catalog, and (D) the input of the answer label. Note that the questions and answers appearing in the illustration are fictional since we cannot publish the actual data due to restrictions by the data provider.

computing an unsupervised similarity between questions, as shown in our empirical evaluation of various methods in Chapter 4.

We implement another version of the illustrated tool that incorporates the SEQUENTIAL annotation approach. This version is identical to the one illustrated in Figure 5.1, with the only difference that *component B*, showing the ranked list of similar questions, is removed. As an expected outcome, the worker using this version of the annotation tool labels each candidate question one by one.

### 5.2.2 Results & Discussion

We evaluate the GROUP-WISE and the SEQUENTIAL annotation approach from the perspective of time efficiency and annotation quality. For that, we employ two workers as human annotators: The first annotator (referred to as *A1*) uses the SEQUENTIAL approach to annotate the 500 customer-support questions. The second annotator (referred to as *A2*) uses the GROUP-WISE approach to annotate the same 500 questions. The annotators in our experiments are students<sup>2</sup> of the Technical University Vienna (TU Wien). To align the annotators' conception on conducting the task, we performed a small-scale test run before starting with the actual annotation run.

To label the 500 customer-support questions using the SEQUENTIAL approach, the first annotator, *A1*, goes through the questions one by one and labels them with the answers selected from the catalog. Questions that could not be answered—e.g., questions where no answer exists in the catalog or questions without an actual information need (e.g., "Hello!", "what's the purpose of life"<sup>3</sup>)—were labeled with *no-answer*. The annotation took a total of 508 minutes with an average time of 61 seconds per question (breaks excluded). From the possible 373 answer entries, 99 answers occurred, and 211 questions were labeled with *no-answer*. Considering that we work with questions asked to a chat-bot system, the high number of *no-answer* questions is not surprising since users often ask the chat-bot irrelevant questions (e.g., "How old are you?") or submit texts that are not actual questions (e.g., "Hi").

The second annotator *A2* labeled the 500 customer-support questions via the GROUP-WISE approach. The annotation took a total of 209 minutes, with an average of 25 seconds per question.

<sup>2</sup>Annotator *A2* being the author of this thesis

<sup>3</sup>All questions are translated from German

The annotator went through 242 candidate questions to label the 500 questions with an answer from the catalog, showing that many candidates were labeled in groups together with other questions. From the possible 373 answer entries, 79 answers occurred, and 200 questions were labeled as *no-answer*. Particular to this annotation approach, even some of the *no-answer* questions were annotated together, such as questions with greeting formulas (e.g., "Hello!", "hi").

The two annotators agreed on the answer label for 362 questions (72%) and disagreed for 138 questions (28%). Such a degree of disagreement can be expected for this annotation task since selecting a suitable answer is often subjective, depending on the annotator's interpretation of the question. For example, consider the question "new phone number", for which the answer catalog contains two potentially relevant answers: one about changing the phone number and another one about requesting a specific phone number. The concrete answer label selected for this question depends on the annotator's subjective interpretation. To give an idea of the questions that are frequently asked, we give an overview of the most frequently assigned answers in Figure 5.2.

### Comparison of the Efficiency

We first compare the time efficiency between both approaches based on the number of answer lookups. Figure 5.3a shows that by using the GROUP-WISE approach, the number of lookups to find relevant answers is reduced by 51% compared to the SEQUENTIAL approach. Notice that for annotating the first 300 questions, less than 100 lookups are performed in the GROUP-WISE approach. Furthermore, notice that from 300 to 500 questions, the number of answer lookups grows similarly for both approaches. The reason for that is that in the end, mostly unique questions remain, which cannot be annotated as a group with other questions. While the reduction of answer lookups is a positive indicator of efficiency, the GROUP-WISE approach also has the overhead of grouping the questions with the same information need. To take this overhead into account, we compare the elapsed working time of the annotators next.

A specific consideration when comparing working time is that it varies depending on the annotator's speed. To reduce this bias, we estimate the annotation time of the worker A2 when using the SEQUENTIAL approach instead of the GROUP-WISE approach. For that, A2 labels a set of 50 questions that are randomly sampled from the pool of the 500 questions. The median time to annotate these questions was 43 seconds per question, resulting in the estimation of 358 minutes to annotate the 500 questions. We use this estimation to compare the elapsed working time between both annotation approaches.

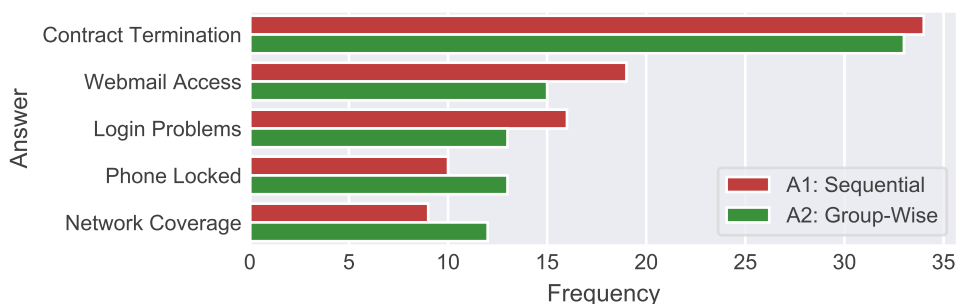


Figure 5.2: The top-5 most frequently assigned answers by the two annotators. For conciseness, we aggregated the full answer text to a few keywords for this plot.

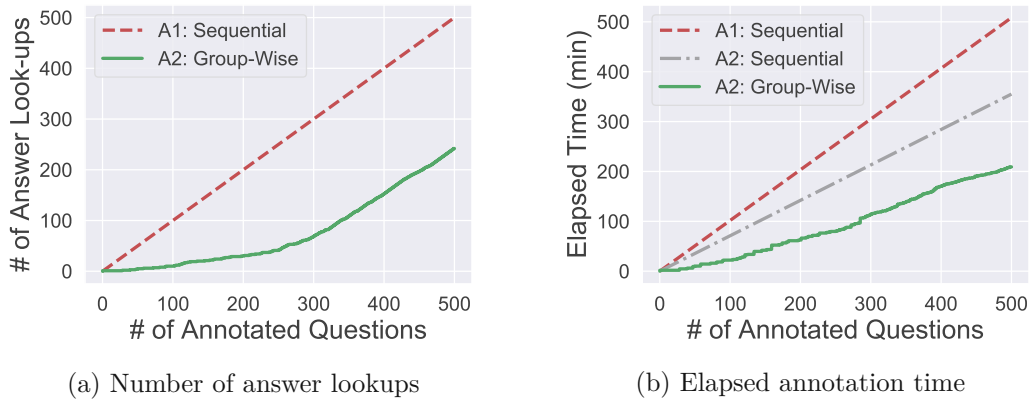


Figure 5.3: Comparison of the efficiency between the GROUP-WISE approach and the SEQUENTIAL approach based on the customer-support dataset

The comparison of the working times is presented in Figure 5.3b. The illustrated line plots for the SEQUENTIAL approach are based on the median annotation times of the annotators *A1* and *A2* (*A2* based on the described estimation). The small spikes in Figure 5.3b particular to the GROUP-WISE approach indicate the overhead for selecting similar questions with respect to the candidate questions. Notice that annotator *A2* needed approximately 358 minutes to annotate the 500 customer-support questions when using the SEQUENTIAL approach and only 209 minutes when using the GROUP-WISE approach – which is a speed-up of 41%. This finding shows that the GROUP-WISE annotation approach is more time-efficient than the SEQUENTIAL approach for labeling the 500 customer-support questions.

### Comparison of the Annotation Quality

So far, we have analyzed the efficiency of the two approaches. Now we investigate whether the GROUP-WISE approach has a negative impact on the annotation quality. For that, we first create a final set of annotations to which we compare the annotations of *A1* and *A2*. The final set consists of answer labels for that both annotators agreed and answer labels for that they disagreed adjudicated by a third TU Wien student employed as a meta-annotator.

Based on this final set of annotations with adjudicated disagreements, we measure the annotators' label quality by computing the inter-annotator agreement via Cohen's Kappa (Section 2.4.1). Between the final set and the set provided by *A1* with the SEQUENTIAL approach, we measure a Kappa agreement of  $\kappa = 0.80$ . Between the final set and the annotations from *A2*, we measure an agreement of  $\kappa = 0.83$ .

The similar agreement of *A1* and *A2* to the final set indicates that both annotators made a similar number of mistakes (or have a similar degree of disagreement to the meta-annotator). Given the high similarity between the Kappa agreement of both approaches, we conclude that the label quality is not affected (positively nor negatively) by using the GROUP-WISE approach to annotate the customer-support questions.

### 5.3 Biomedical Named-Entity Annotation

The second data source on which we apply the GROUP-WISE annotation approach is case study reports. A case study report is a publication describing the course of a patient having a certain medical condition. Case study reports describe several characteristics of a patient, such as age, gender, symptoms, conducted laboratory tests, or pre-existing conditions. Medical practitioners write and publish the reports to make the obtained knowledge available to peers who can learn from the described cases. The availability of an annotated corpus of case reports is a beneficial resource for many computer-assisted medical tasks, such as diagnosis [NFFZ17], the characterization of rare diseases [DPL<sup>+</sup>13], or the discovery of unexpected associations between diseases and symptoms [ZMBS14, dVGS<sup>+</sup>18].

In this section, we describe our experiments for a time-efficient annotation of case study reports. Specifically, we ask human annotators from the Mechanical Turk platform to label the named-entities *Age*, *Gender*, and *Symptom* in the text of medical case study reports. For this, we first split the reports into individual sentences, which we then present to the crowdworkers for annotation. By manually examining the sentences of case reports, we found that a similar wording and phrasing is often used across different reports. The similar wording and phrasing means that sentences often have a similar textual structure, as illustrated in Table 5.1a.

We apply the GROUP-WISE approach to improve the time efficiency of the annotation procedure. We bundle semantically similar sentences into groups of three and present the bundled sentences to the workers for annotation. The bundling of samples allows workers to process similar sentences time-efficiently. In particular, we show through our experiments that workers are more time-efficient for two critical aspects: (i) The overhead to switch between the different named-entities labels is reduced, and (ii) workers process the similar sentences quickly, reducing the average working time needed per sentence.

We illustrate the advantage of bundling semantically similar sentences in Table 5.1. The sentences bundled in Table 5.1a show the advantage in using the GROUP-WISE approach, as each sentence requires a similar annotation at almost the same positions in the text. Therefore, a worker can label multiple entities across the three sentences with only a few label selections, which reduces the overall overhead for switching between labels. As an additional benefit in presenting similar sentences, the cognitive effort for processing the similar sentences is reduced. The cognitive effort is usually more demanding in the SEQUENTIAL approach since the sentences are randomly selected, and therefore, they are often not similar to each other, as shown in Table 5.1b.

#### 5.3.1 Experiment Setup

This section describes our experiment setup for annotating case study reports using the GROUP-WISE annotation approach. We describe the used data source, the annotation tool, and the annotators recruited for our experiment.

#### Dataset & Preparation

The case reports that we annotate in our experiment were published in the *BMJ Case Reports* journal<sup>4</sup>. This journal contains over 20 thousand reports from all medical disciplines. For a case report to be accepted to the BMJ Case Reports journal, the case must provide grounds for

<sup>4</sup><https://casereports.bmj.com/>



Table 5.1: Sentences from case study reports with annotations for Age, Gender, and Symptom.

(a) Three similar sentences as they appear in the GROUP-WISE approach

---

A 68-year-old lady with end-stage chronic obstructive pulmonary disease presented with vomiting and abdominal pain .

---

A 68-year-old man with expressive dysphasia presented with upper gastrointestinal haemorrhage , jaundice and abdominal pain .

---

A 20-year-old Japanese woman emergently presented with the chief complaint of pain at the right iliac fossa .

---

(b) Three randomly sampled sentences as they appear in the SEQUENTIAL approach

---

A 68-year-old lady with end-stage chronic obstructive pulmonary disease presented with vomiting and abdominal pain .

---

Ultrasound and CT scans revealed an acalculous grossly thickened gallbladder , with high attenuation non-echogenic material both within and surrounding the structure .

---

This atypical presentation , along with the unusual FVL , led to a significant delay in the diagnosis of the tracheal mass .

---

discussion and a rich learning experience for other medical practitioners reading the report. The medical conditions that are described in the cases can be rare or common.

In our experiment, we annotate 90 sentences extracted from case reports of the BMJ Case Reports journal. We sampled the 90 sentences as follows: From the over 20 thousand reports, we filtered for reports that describe cases of the diseases: COVID-19, Sleep Apnea, Migraine, Cholecystitis, Measles, and Chronic Obstructive Airway Disease. In total, we found 227 case reports about these diseases. We segmented the sentences of these reports using the CoreNLP library [MSB<sup>+</sup>14] and selected 90 sentences for our evaluation randomly.

We prepare the 90 sentences for annotation on the Mechanical Turk crowdsourcing platform. We prepare the sentences for the GROUP-WISE annotation approach following the procedure of Section 5.1 as follows: First, a candidate sentence is randomly selected from all the available sentences. Second, the two most similar sentences to the candidate are retrieved via an unsupervised semantic similarity method. We computed the semantic similarity between sentences using Sent2Vec with the BioSent2Vec model, which we selected based on our evaluation of various methods in Chapter 4. Third, the randomly selected sentence and the two most similar sentences are bundled. This bundle represents one HIT on the Mechanical Turk platform. We repeat the described steps until all 90 sentences are bundled, generating a total of 30 HITs. For the SEQUENTIAL approach, we also bundle the sentences into groups of three, but this time, the sentences are bundled randomly, without considering their semantic similarity.

Figure 5.4: The annotation interface for the Mechanical Turk platform for labeling Age, Gender, and Symptom in case study reports

### Annotation Tool

We developed an annotation interface compatible with the Mechanical Turk platform. The interface is illustrated in Figure 5.4 and offers the following functionality:

- Workers can consult the annotation instructions through an expansible/collapsible panel, shown at the top of Figure 5.4. The instructions are available in Appendix A.5.
- Workers can switch between the labels Age, Gender, and Symptom.
- The worker can highlight text in the three shown sentences by clicking and dragging over the relevant text. Upon mouse release, the dragged-over text will be annotated with the currently selected label. Clicking on an annotated text removes the annotation.
- To submit the highlighted text and finalize the current HIT, the worker presses the *Submit* button. In case no relevant information occurs for Age, Gender, and Symptom, the worker needs to mark the corresponding checkbox as a confirmation.

The illustrated annotation interface is the same for workers of the SEQUENTIAL and the GROUP-WISE approach. The only difference is in the earlier described sample presentation, where the three sentences are either randomly sampled (SEQUENTIAL approach) or semantically similar to each other (GROUP-WISE approach). Notice that the three sentences illustrated in Figure 5.4 were generated for the GROUP-WISE annotation approach, recognizable by their semantic similarity.

### Annotators

We recruited annotators from the Mechanical Turk platform using the following criteria: Each crowdworker needed at least 500 accepted HITs and an acceptance rate of at least 95% on previous tasks. We conducted a small-scale test run to identify workers that are good at the task. In the test run, each worker had to annotate 10 sentences. Workers reaching at least a 70% Kappa agreement for labeling Age/Gender and a 50% agreement for the more difficult entity of Symptom were included in the full-scale run. Rather than dividing the qualified workers randomly into



using the GROUP-WISE or SEQUENTIAL approach, we divided them more fairly based on their working time to complete the test run. For example, the quickest worker was assigned to use the SEQUENTIAL approach, the second quickest to use the GROUP-WISE approach, and so on. By equally dividing workers based on their working time during the test run, we reduce the chance of a biased assignment of workers to one of the two approaches. In total, 27 workers qualified through the test run. Of these, 14 were assigned to use the GROUP-WISE approach and 13 to use the SEQUENTIAL approach. Workers were not informed about the experimental nature of this task, as this could be another potential bias.

### 5.3.2 Results & Discussion

We use the 90 sentences extracted from case study reports to compare the efficiency between the SEQUENTIAL and the GROUP-WISE annotation approach. Each sentence is labeled by five crowdworkers to generate sufficient data for our evaluation. Since the 90 sentences were grouped into bundles containing three sentences each, this makes a total of  $30 \times 5 = 150$  HITs per approach and 300 HITs in total. Based on these annotated HITs, we will now analyze the two annotation approaches with respect to time efficiency and label quality.

#### Comparison of the Efficiency

We first compare the time efficiency based on the working time that was needed to complete a single HIT. The working time per HIT is automatically recorded by the Mechanical Turk platform as a built-in feature and is the duration that elapses from acceptance until submission of a HIT. The results for the working time of both approaches are presented in Table 5.2. The table shows that the total working time to annotate the 150 HITs was  $\approx 14$  hours in the SEQUENTIAL approach, compared to  $\approx 10$  hours in the GROUP-WISE approach – which is an improvement in time efficiency of nearly 29%.

As an additional measurement of time efficiency, we record the number of clicks needed to switch between the three labels Age, Gender, and Symptom. Fewer clicks are better since the number of clicks indicates the overhead for workers to interact with the annotation tool. Therefore, keeping the interaction overhead as small as possible is desirable [DKC<sup>+</sup>18].

We present the total number of clicks of workers of the GROUP-WISE and the SEQUENTIAL approach in Table 5.3. The table shows that workers using the GROUP-WISE approach conducted fewer clicks and, therefore, fewer interactions for annotating the 150 HITs. In total, we found

Table 5.2: Comparison of the time efficiency between SEQUENTIAL and GROUP-WISE approach for annotating case study reports. The times are recorded during annotating 450 sentences (= 150 HITs) by once using the SEQUENTIAL and once the GROUP-WISE approach. We rounded the hours and seconds up to the next integer value.

|                      | Working Time |             |
|----------------------|--------------|-------------|
|                      | SEQUENTIAL   | GROUP-WISE  |
| Total                | 14 hours     | 10 hours    |
| Average per HIT      | 343 seconds  | 242 seconds |
| Average per sentence | 114 seconds  | 80 seconds  |

Table 5.3: The number of clicks for crowdworkers switching to the labels Age, Gender, and Symptom. The presented numbers are recorded over annotating the 450 sentences (= 150 HITs) by once using the SEQUENTIAL and once the GROUP-WISE approach. Fewer clicks indicate a reduced overhead for switching between the labels.

|         | #Clicks    |            |
|---------|------------|------------|
|         | SEQUENTIAL | GROUP-WISE |
| Age     | 46         | 48         |
| Gender  | 186        | 139        |
| Symptom | 185        | 165        |
| All     | 417        | 352        |

that workers using the GROUP-WISE required 16% fewer clicks to switch between the labels than workers of the SEQUENTIAL approach. This is as expected since workers of the GROUP-WISE approach can often re-use a selected label to annotate named-entities across the three presented sentences. On the other hand, in the SEQUENTIAL approach, there is a higher chance that workers select a label to annotate a single sentence without the possibility to re-use the same label for the other sentences. Notice that no substantial difference in the number of clicks between the GROUP-WISE and the SEQUENTIAL approach was found for the label Age. This is because the label Age is pre-selected as the default label and annotators of both approaches rarely need to switch to this label.

We analyzed the number of clicks aggregated over the entire worker base rather than examining individual workers. We do not provide results for individual workers since the Mechanical Turk platform allows workers to decide on a case-to-case basis whether they would like to work on a specific HIT or not. As a result, the task was completed in a collaborative effort by the 27 crowdworkers, where each worker participated to a different degree, making a meaningful analysis of individual workers difficult.

### Comparison of the Annotation Quality

Next, we compare the label quality for annotations collected by workers of both approaches. For this, we aggregate the five individual annotations of the GROUP-WISE and the SEQUENTIAL approach via a majority voting on a token-level. Afterward, we measure the label quality between the final set of annotations and a set of expert annotations by computing the inter-annotator agreement via Cohen’s Kappa. The expert annotations are created by a medical doctor working in a hospital’s intensive care department. This doctor is fluent with the jargon prevailing in case study reports, has profound expertise in disease-symptom diagnosis, and has experience with manual text annotation.

The comparison of Kappa agreements between the non-expert and the expert annotations is presented in Table 5.4. The table shows nearly perfect agreements for labeling Age and Gender. These two labels are the easiest to annotate since they do not require medical expertise to be labeled correctly. The slight disagreement for the label Gender is caused by crowdworkers missing personal pronouns that indicate the gender of a patient (e.g., *her*, *him*, *she*). We explicitly asked in the annotation instructions to also label these pronouns. A slightly lower agreement was observed for the label Symptom. Symptoms are more difficult to label, especially since we

Table 5.4: Cohen’s Kappa agreement between expert annotations and aggregated crowd-worker annotations of both approaches.

|         | Cohen’s Kappa ( $\kappa$ ) |            |
|---------|----------------------------|------------|
|         | SEQUENTIAL                 | GROUP-WISE |
| Age     | 1.00                       | 1.00       |
| Gender  | 0.98                       | 0.98       |
| Symptom | 0.82                       | 0.83       |

asked annotators to include all characteristics of a symptom, such as severity, duration, and localization. To accurately do so, workers needed to identify and label complex phrases, such as "severe abdominal pain in the upper right quadrant of the stomach". Another difficulty for workers was the differentiation between symptoms and diseases, which is not always distinct. For example, one could argue that "chronic coughing" is a disease, but at the same time, it could also be a symptom of another disease such as chronic airway obstruction. Overall, the results show a similar agreement between the expert annotations and the aggregated annotations of either approach. This result suggests that using the GROUP-WISE annotation approach does not harm the crowdworkers’ ability to assign labels accurately.

## 5.4 Summary

We proposed the GROUP-WISE annotation approach to create new corpora for IR and NLP-related tasks time-efficiently. The idea behind the GROUP-WISE approach is that groups of semantically similar samples can be annotated quicker than labeling each sample individually, as is the case in the SEQUENTIAL approach.

We showed that the GROUP-WISE approach is especially time-efficient for creating annotations for the task question-answering. For this task, obtaining the correct answer label for a question is usually time-intensive; however, by applying the GROUP-WISE approach, we showed that only a single answer lookup is often sufficient to label large groups of questions. Similarly time-efficient was the GROUP-WISE approach for labeling named-entities in case study reports. Also for this task, we found that grouping similar samples allows workers to process the samples more time-efficiently and reduces the workers’ overhead for interacting with the annotation interface.

We compared the GROUP-WISE approach to the SEQUENTIAL approach from the two perspectives of time efficiency and label quality. As a measure for label quality, we computed the Cohen’s Kappa inter-annotator agreement between workers of both approaches and a set of gold standard annotations. As a measure for time efficiency, we computed the raw working time and other efficiency-related metrics, such as the number of clicks and answer lookups. We summarize our results for the conducted tasks question-answering and named-entity annotation next.

### 5.4.1 Question Answer Annotation

The first task for which we evaluated the GROUP-WISE annotation approach was question-answering in the customer-support domain. Two annotators were employed: the first used the SEQUENTIAL approach, and the second used the GROUP-WISE approach. Both annotators conducted the same task of labeling 500 questions with an answer label from a catalog containing

373 entries. The worker of the GROUP-WISE approach was presented with candidate questions and a ranked list of similar questions per candidate. The worker skimmed through the ranked list and selected questions with the same intent as the candidate question. Finally, the worker labeled the candidate question and the selected questions by conducting only a single lookup to find the appropriate answer label. The worker of the SEQUENTIAL approach processed the questions one by one and assigned an answer label to each question individually.

We compared the GROUP-WISE approach to the SEQUENTIAL approach with respect to the number of answer lookups, annotation time, and label quality. We showed that the GROUP-WISE approach retains the same label quality, is 41% more time-efficient, and requires 51% fewer answer lookups. The reduction of answer lookups shows that the worker of the GROUP-WISE approach spent substantially less time searching for the correct answer within the 373-entry catalog. We conclude that the GROUP-WISE approach can be used as a time-efficient alternative to the SEQUENTIAL approach for the task of question-answer annotation.

### 5.4.2 Biomedical NER Annotation

We further applied the GROUP-WISE approach to the task of named-entity annotation in medical case study reports. We recruited workers from the Mechanical Turk platform to label the named-entities Age, Gender, and Symptom. The data annotated consisted of 90 sentences extracted from case study reports. The sentences were presented in bundles containing three sentences each to the crowdworkers. The bundles for the GROUP-WISE approach contained semantically similar sentences, and the bundles for the SEQUENTIAL approach were randomly sampled.

We compared the GROUP-WISE approach to the SEQUENTIAL approach with respect to annotation time, label quality, and the number of clicks for switching between the named-entity labels. The number of clicks indicates the overhead for workers to switch between the different labels and is an indirect measurement of time efficiency. We found that crowdworkers using the GROUP-WISE approach are 29% quicker and require 16% fewer label switches than workers of the SEQUENTIAL approach. The evaluation of label quality showed no substantial difference in labeling Age, Gender, and Symptom. The conducted annotation of named-entities with crowdworkers is the proof-of-concept for combining the GROUP-WISE approach with crowdsourcing platforms.

### 5.4.3 Resources

We publish the annotated corpus and the annotation interface for the named-entity labeling task on our GitHub page<sup>5</sup>. Unfortunately, we cannot release the data or annotations for the question-answering task because of restrictions by the data provider.

---

<sup>5</sup><https://github.com/Markus-Zlabinger/casereports>

# Supporting Non-Experts for Complex Annotation Tasks

The success of crowdsourcing-based annotation of text corpora depends on ensuring that crowdworkers are sufficiently well-trained to perform the annotation task accurately. Reaching a certain quality threshold is challenging, especially for tasks that require specific expertise to be performed (e.g., tasks of the medical domain [NLP<sup>+</sup>18]).

The common approach to compensate for the missing knowledge of individual non-expert workers is to train them via task instructions and a few example cases that demonstrate how the task should be performed [NLP<sup>+</sup>18, SOJN08]. These globally defined *task-level* examples, however, often (i) only cover the common cases that are encountered during an annotation task and (ii) require effort from crowdworkers during the annotation process to find the most relevant example for the currently annotated sample.

In this chapter, we address these limitations with a new annotation approach called Dynamic EXamples for Annotation (DEXA). In addition to task-level examples, annotators are supported with *dynamic examples* that are semantically similar to the currently annotated text sample. The dynamic examples are retrieved from data samples previously annotated by experts. Such expert samples are usually available since they are crucial to measure the quality of non-expert annotators [SOJN08, DKC<sup>+</sup>18, DKBH16]. We propose to split the expert samples into training samples from which dynamic examples are retrieved and test samples injected into the annotation process to measure worker performance.

We apply the DEXA approach to a task of the medical domain, known as the PIO<sup>1</sup> task. In the PIO annotation task [HLD06, NLP<sup>+</sup>18], annotations are collected for the *Participants* (e.g., patients with headache), *Interventions* (e.g., Ibuprofen), and *Outcomes* (e.g., pain reduction) of clinical trial reports. To perform the PIO task accurately, annotators usually require fundamental medical expertise to understand the terminology and jargon prevailing in the biomedical literature. We compensate for the lack of medical expertise by supporting non-expert annotators via the DEXA approach. Our results show that non-expert annotators supported by the DEXA approach reach high inter-annotator agreements to experts with an average of 0.78/0.75/0.69 for P/I/O.

<sup>1</sup>The difference to the PICO task is that Intervention/Control are not differentiated [NLP<sup>+</sup>18]

The contributions of this chapter are as follows:

- We propose the DEXA annotation approach for collecting high-quality annotations from non-experts.
- We apply the approach to the complex PIO annotation task and show high agreements between experts and non-experts supported by the DEXA approach.

The remainder of this chapter is structured as follows: We define the DEXA annotation approach in Section 6.1. We describe the annotation task of labeling Participants, Interventions, and Outcomes in clinical trial reports in Section 6.2. The experiment setup of applying the DEXA approach to the PIO task is described in Section 6.3, followed by the presentation and discussion of the results in Section 6.4. The chapter is summarized in Section 6.5.

## 6.1 Dynamic Examples for Annotation

In this section, we formally define the Dynamic EXamples for Annotation (DEXA) approach. The approach consists of showing examples to annotators on a *task-instance level*<sup>2</sup>, i.e., dynamic to the currently annotated text sample. Given a set of labeled expert samples  $E$  and a set of samples  $U$  to be labeled by non-experts, the DEXA annotation approach consists of the following steps:

1. The samples of  $E$  are divided into a test set  $E_{te} \subset E$  and a training set  $E_{tr} \subset E$ , where  $E_{te} \cap E_{tr} = \emptyset$ . From the training set, the dynamic examples are drawn. The samples from the test set are injected into  $U$  to measure the quality of the non-expert annotators, resulting in the annotation set  $A = U \cup E_{te}$ .
2. An unsupervised similarity method  $\text{sim}(p, a) \in \mathbb{R}$  is selected to compute the semantic similarity between a sample  $p \in E_{tr}$  of the training set to a sample  $a \in A$  of the annotation set. The similarity method should be selected based on the task at hand. For example, in our experiments, samples are sentences, and therefore, we use an unsupervised semantic sentence-to-sentence similarity method.
3. The annotation set  $A$  is labeled by non-experts. For each unlabeled sample  $a$ , the similarity method  $\text{sim}(p, a)$  is used to compute the similarity to each sample in the training set, i.e.,  $\text{sim}(p_1, a), \dots, \text{sim}(p_n, a) \forall p \in E_{tr}$ . Then, the top  $k$  most similar samples  $p_1, \dots, p_k \in E_{tr}$  are shown as dynamic demonstration examples to the annotators.
4. Finally, the accuracy of non-expert annotators is compared to that of expert annotators based on the test samples  $E_{te}$  that were injected into the annotation set  $A$  in *Step 1*.

We evaluate the DEXA approach based on the task of labeling Participants, Interventions, and Outcomes in clinical trial reports.

## 6.2 PIO Annotation Task

Evidence-Based Medicine is the practice of decision-making based on the best available scientific information [SRG<sup>+</sup>96]. Finding such information rapidly is essential, especially in the current pandemic crisis, where thousands of medical articles about COVID-19 are published weekly [ŠGP20].

<sup>2</sup>A *task-instance* is known as a Human Intelligence Task (HIT) on the Mechanical Turk platform.

To make the search process time-efficient, the PICO model enables specific search for: Participants (e.g., patients with headache), Interventions (e.g., Ibuprofen), Comparisons (e.g., placebo), and Outcomes (e.g., pain reduction) [HLD06]. To allow a search for structured PICO information in trial reports, a prior automatic extraction is necessary.

The effectiveness of an automatic PICO extraction depends on the quality of manually annotated corpora. As an alternative to scarce and expensive expert annotators, Nye et al. [NLP<sup>+</sup>18] hired crowdworkers from the Mechanical Turk platform (MTurk) to annotate Participants, Interventions, and Outcomes in clinical trial reports. The crowdworkers, however, reached low inter-annotator agreements to experts, potentially affected by (i) a lack of domain-specific expertise of the crowdworkers and (ii) an uneven task length distribution.

- **The lack of domain-specific expertise** makes it difficult for crowdworkers to understand the terminology and jargon that prevails in the medical literature [KMCY11, Wal18a]. As a result, workers experience medical tasks as cognitively overwhelming, with the side effect of a decreased label quality [FKTC13].
- **An uneven task length distribution** makes the effort to complete individual task-instances unevenly distributed, thus enticing workers to "cherry-pick" short samples or rush longer ones [CTIB15, FSL<sup>+</sup>18]. In the task design by Nye et al. [NLP<sup>+</sup>18], *entire abstracts* of clinical trial reports were annotated. These abstracts contain, on average, 268 words with a high standard deviation of 89, resulting in an uneven task length distribution.

We propose two novel annotation approaches to address these problems: SENBASE and SENSUPPORT. We systematically compare the two approaches to the approach by [NLP<sup>+</sup>18], referred to as the BASELINE approach. The three approaches are described in detail next.

### 6.2.1 Baseline

In the task design of [NLP<sup>+</sup>18], entire abstracts of clinical trial reports are presented to annotators who are asked to label the PIO entities. The annotation of Participant, Intervention, and Outcome is conducted as three individual sub-tasks to reduce the cognitive overload needed to switch between the labels. For each sub-task, annotation guidelines are crafted to prepare the workers. The guidelines consist of a few *static examples*, which illustrate how the task should be performed, and annotation instructions, which describe what text spans should or should not be annotated as PIO.

### 6.2.2 SenBase

The annotation of entire abstracts leads to an uneven distribution of task effort to complete individual task-instances. We illustrate this problem in Table 6.1, where we compare the word counts of abstracts to sentences. The table shows that the annotation of sentences leads to a better distribution in task effort, indicated by the substantially lower standard deviation of 13 compared to abstracts with a standard deviation of 89.

Based on this analysis, we propose a new task design, SENBASE, in which we switch from abstract to sentence annotation. Specifically, we split each abstract into individual sentences, in which annotators label the PIO entities – or mark a checkbox if no PIO entity could be identified. Similar to the BASELINE, the task is divided into three individual sub-tasks for PIO, and the annotators are trained with a few examples and instructions. The examples and instructions conform to the annotation guidelines used by [NLP<sup>+</sup>18] and are available in Appendix A.6.



Although the annotation of sentences improves the distribution in task effort, sentences might appear out-of-context. This means that two consecutively annotated sentences could stem from two different abstracts. The inability to preserve a certain order for presenting text samples is typical for crowdsourcing platforms since workers can usually (i) skip individual task-instances and (ii) start/stop working on task-instances arbitrarily. The lack of context can be problematic since the context is essential, e.g., to identify the meaning of an abbreviation that was defined in an earlier sentence. To address the lack of context, we give workers access to the entire abstract via an expandable window.

### 6.2.3 SenSupport

This approach extends the SENBASE approach by additionally addressing the lack of domain-specific expertise of crowdworkers. The common approach to train crowdworkers for difficult tasks is to provide a few examples illustrating how the task should be performed. Providing examples is essential for a successful task design [DKC<sup>+</sup>18]; however, examples are usually defined *statically* over an entire task and might not be helpful at individual text samples. To improve the effectiveness of examples, we propose the SENSUPPORT task design incorporating the proposed DEXA approach to support annotators. The DEXA approach supports the annotators with *dynamic examples* specific to the currently annotated sentence. We retrieve dynamic examples from a small set of sentences previously annotated by experts. Aligned with the SENBASE and BASELINE approach, we divide the task into three sub-tasks for PIO and provide the same instructions and examples. To address the lack of context of sentences, we provide access to the entire abstract via an expandable window, as in the SENBASE approach.

We illustrate a few dynamic examples for the task of PIO annotation in Table 6.2. The first three cases *a)*, *b)*, and *c)* show dynamic examples for annotating Participant, Intervention, and Outcome, respectively. The fourth case *d)* shows that the dynamic example provides strong support in annotating all PIO entities – even though the sentence is rather complex and long. Finally, the last case *e)* shows that no appropriate dynamic example is found for the sentence. In such cases, workers need to decide independently.

The dynamic examples that we show to support annotators are retrieved using an unsupervised semantic short-text similarity method. In our experiments, we retrieve similar sentences via the sentence embedding method BioSent2Vec [CPL19]. We selected this method based on our evaluation of ten methods in Chapter 4. We found that BioSent2Vec is an optimal choice since (i) it is the state-of-the-art for various short-text similarity tasks in the biomedical domain, and

Table 6.1: Analysis of the word counts of abstracts versus sentences. We measure the word count based on tokenized text, excluding punctuation. The data basis of this analysis is the EBM-NLP corpus, described in Section 6.3.

|          | # Words |      |      |        |
|----------|---------|------|------|--------|
|          | Min.    | Max. | Avg. | Stdev. |
| Abstract | 57      | 562  | 268  | 89     |
| Sentence | 5       | 105  | 25   | 13     |



Table 6.2: Text samples with dynamic examples for **Participants**, **Interventions**, and **Outcomes**. Note that only the labels for either P, I, or O (depending on the sub-task) within the dynamic examples are visible to workers. The labels shown in the text samples should be highlighted by the workers.

|    |                 |  |
|----|-----------------|--|
| a) | Text Sample     | <u>Thirty-nine subjects</u> completed the study and were included in the data analysis.  |
|    | Dynamic Example | <u>Ninety-three subjects</u> were randomly assigned.   |
| b) | Text Sample     | <u>QYJDR</u> is an effective formula in treatment of EMs-related infertility.  |
|    | Dynamic Example | <u>Eltrombopag</u> is an oral thrombopoietin receptor agonist for the treatment of thrombocytopenia.   |
| c) | Text Sample     | There were no serious <u>adverse events</u> .  |
|    | Dynamic Example | <u>Adverse events</u> did not significantly differ in the 2 groups.  |
| d) | Text Sample     | We performed a randomized, controlled study comparing the <u>prophylactic effects</u> of capsule forms of <u>fluconazole</u> (n=110) and <u>itraconazole</u> (n=108) in <u>patients with acute myeloid leukemia (AML) or myelodysplastic syndromes (MDS) during and after chemotherapy</u> . |
|    | Dynamic Example | A randomized, double-blind, placebo-controlled study on the <u>immediate clinical and microbiological efficacy</u> of <u>doxycycline</u> (100mg for 14 days) was carried out to determine the benefit of adjunctive medication in <u>16 patients with localized juvenile periodontitis</u> . |
| e) | Text Sample     | The majority (63%) of the project group had no <u>admission</u> during the 10 month study period.  |
|    | Dynamic Example | Referral occurred at any stage of the patients' EECU admission.  |

(ii) a pre-trained model is available<sup>3</sup> trained on PubMed [CPL19], which is the same underlying data source as the clinical trial reports used in our experiments.

## 6.3 Experiment Setup

As data source for our experiments, we consider the 191 clinical trial reports of the EBM-NLP corpus [NLP<sup>+</sup>18]. For each trial report in this corpus, gold standard labels for PIO are available assigned by medical expert annotators. The reports originate from PubMed and consist of a title and an abstract. As preprocessing steps, we use the Stanford CoreNLP library [MSB<sup>+</sup>14] to segment and the NLTK library [BKL09] to tokenize the sentences. Afterward, we randomly split the 191 reports into a test set  $E_{te}$  (41 reports with 423 sentences) used for evaluation and a training set  $E_{tr}$  (150 reports with 1,636 sentences) used to retrieve dynamic examples for the SENSUPPORT approach. Note that the test sentences are usually injected into a much larger set  $U$  for which no gold labels are available (see *Step 1* in Section 6.1); however, in our experiment, we aim to evaluate different annotation approaches and therefore only sentences are annotated that overlap with the gold standard.

<sup>3</sup><https://github.com/ncbi-nlp/BioSentVec>

The sentences of the training set are used for the retrieval of dynamic examples for the SENSUPPORT approach. We retrieve the top-3 most similar sentences for each sentence in the test set and show them as dynamic examples to the crowdworkers.

The samples of the test set are used to compare the three annotation approaches. The annotations for the BASELINE approach are downloaded from the corresponding GitHub page<sup>4</sup>, published in the scope of [NLP<sup>+</sup>18]. The annotations for SENBASE and SENSUPPORT are specifically collected for this study. For the collection, we follow the same annotation setup of [NLP<sup>+</sup>18], namely: annotations are acquired from crowdworkers of the Mechanical Turk platform; workers require a minimum approval rate of 90% on previous tasks to participate; spammers are removed in a small-scale test run; and finally, the payment per HIT is set to \$0.06 per sentence (which we reduced from \$0.30 to reflect the reduced effort needed to complete a HIT).

To collect the annotations for SENBASE and SENSUPPORT, we develop two annotation interfaces for the Mechanical Turk platform. The first interface, used by workers of the SENBASE approach, is illustrated in Figure 6.1a. The second interface, used by workers of the SENSUPPORT approach, is shown in Figure 6.1b. In both interfaces, workers have the optional choice to read the full abstract text in which the currently annotated sentence appears. In this abstract, the currently annotated sentence is highlighted by a blue border, as shown in Figure 6.2. Notice that both interfaces have a similar design to mitigate a potential bias of annotators interacting with different interfaces. The core difference between the two interfaces is the presentation of the three dynamic examples, which are only shown to workers of the SENSUPPORT approach.

We give an overview of the annotation sets evaluated in our experiments in Table 6.3. For SENBASE and SENSUPPORT, we collect 3 redundant annotations per sentence, resulting in  $423 \times 3 = 1,269$  HITs in each PIO sub-task. In the BASELINE, more redundant annotations were collected of 8-17 (average 11 with a standard deviation of 1.7), explaining the larger number of unique workers compared to SENBASE and SENSUPPORT.

## 6.4 Results and Discussion

We report and discuss the evaluation results for the three annotation approaches. We analyze individual workers, aggregated annotations from multiple workers, and worker feedback on the usefulness of dynamic examples. Furthermore, we conduct a thorough error analysis by studying different types of disagreements that commonly occur between experts and crowdworkers.

<sup>4</sup><https://github.com/bepnye/EBM-NLP>

Table 6.3: Overview of the compared annotation sets

| Design     | #Workers | #Redundant | HIT      |
|------------|----------|------------|----------|
| BASELINE   | 403      | 8 - 17     | abstract |
| SENBASE    | 38       | 3          | sentence |
| SENSUPPORT | 31       | 3          | sentence |

**Study Report** (Click to expand)

**Sentence** (Highlight information about participants by clicking on a start and end word)

A multi-component social skills intervention for children with Asperger syndrome : the Junior Detective Training Program .

Sentence does not contain information about participants

Submit

(a) Annotation interface of SENBASE

**Study Report** (Click to expand)

**Sentence** (Highlight information about participants by clicking on a start and end word)

A multi-component social skills intervention for children with Asperger syndrome : the Junior Detective Training Program .

Sentence does not contain information about participants

**Examples** (Caution: The shown examples might contain missing highlights.)

Social skills training ( SST ) is a common intervention for children with autism spectrum disorders ( ASDs ) to improve their social and communication skills .

Teaching emotion recognition skills to young children with autism : a randomised controlled trial of an emotion training programme .

A randomized controlled study of a social skills training for preadolescent children with autism spectrum disorders : generalization of skills by training parents and teachers ?

Submit

(b) Annotation interface of SENSUPPORT

Figure 6.1: Annotation interfaces developed for the Mechanical Turk platform

**Study Report** (Click to expand)

**Title:** A multi-component social skills intervention for children with Asperger syndrome: the Junior Detective Training Program.

**Abstract:** BACKGROUND The study aimed to investigate the effectiveness of a new multi-component social skills intervention for children with Asperger syndrome (AS): The Junior Detective Training Program. This 7-week program included a computer game, small group sessions, parent training sessions and teacher handouts.  
METHOD Forty-nine children with AS were recruited to participate and randomly assigned to intervention (n = 26) or wait-list control (n = 23) conditions.  
RESULTS Relative to children in the wait-list group, program participants showed greater improvements in social skills over the course of the intervention, as indicated by parent-report measures. Teacher-report data also confirmed that children receiving the intervention made significant improvements in social functioning from pre- to post-treatment. Treatment group participants were better able to suggest appropriate emotion-management strategies for story characters at post-intervention than at pre-intervention, whereas control participants were not. However, there was no difference in the improvements made by children in the intervention and control conditions on facial expression and body-posture recognition measures. Follow-up data suggested that treatment gains were maintained by children at 5-months post-intervention.  
CONCLUSIONS The Junior Detective Training Program appeared to be effective in enhancing the social skills and emotional understanding of children with AS. Limitations and suggestions for future research are discussed.

Figure 6.2: Annotators can examine the entire abstract text through an expandible window in the sentence-based approaches. In the illustrated example, the currently annotated sentence is the title of the abstract, indicated by the blue border.

### 6.4.1 Agreement of Individual Crowdworkers

We measure the label quality between individual crowdworkers and the gold standard annotations by computing the inter-annotator agreement in terms of Cohen’s Kappa (Section 2.4.1). The results in Figure 6.3 show a clear improvement of Kappa scores of the sentence-based task designs compared to the abstract-based task design BASELINE. Substantially higher agreements are reached for labeling Intervention and Outcome. Notable is the outlier of the SENSUPPORT approach for the annotation of Intervention, denoted by a dot. This worker reached a distinctly lower agreement to the gold standard than the other workers of the SENSUPPORT approach.

The results of SENBASE compared to SENSUPPORT show that the utilization of dynamic examples further increases the Kappa agreement, especially for the annotation of Intervention. This additional improvement was obtained without additional monetary compensation to crowdworkers since we pay \$0.06 per HIT in both sentence-based approaches.

The analysis of individual workers has the disadvantage that workers who labeled only a few task-instances are less reliable than workers who labeled several samples. We addressed this problem by limiting the presented analysis to workers who labeled at least 5% of the test set. All workers are considered in the analysis of aggregated annotations, described next.

### 6.4.2 Agreement of Aggregated Annotations

Here, we analyze the label quality of meta-annotations that are aggregated from multiple redundant annotations. We consider two aggregation methods: (i) majority voting (MV), where individual workers are weighted equally, and (ii) Dawid-Skene<sup>5</sup> (DS), where the reliability of individual workers is automatically computed and used for a weighted aggregation [DS79].

We measure the quality of aggregated annotations by computing the Kappa agreement to the gold standard annotations. We compute the aggregations for the sentence-based approaches based on the 3 available redundant annotations. Since there are 8-17 redundant annotations available for the BASELINE approach, we (i) select 3 random annotations, (ii) aggregate them, and (iii) compute the agreement to the gold standard. Since the random selection in (i) can be affected

<sup>5</sup>We use the implementation from [https://github.com/dallascard/dawid\\_skene](https://github.com/dallascard/dawid_skene)

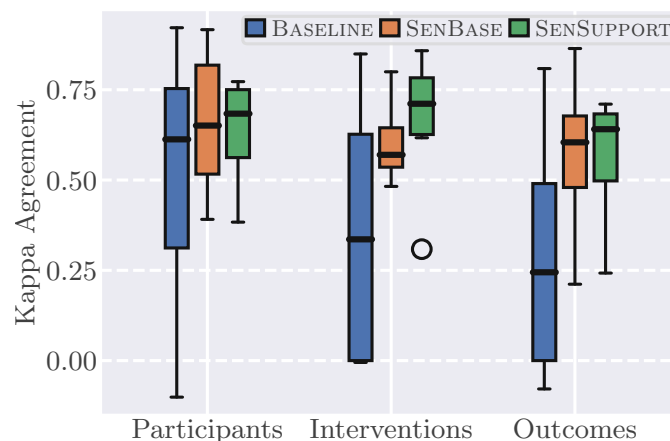


Figure 6.3: Kappa agreements between individual crowdworkers and the gold standard.

by a lucky/unlucky seed, we repeat (i-iii) 20 times and compute a robust final agreement by averaging the 20 individual Kappa scores.

The results for 3 aggregated annotations show that the highest agreements to the gold standard are reached by the SENSUPPORT approach, followed by SENBASE (Table 6.4). Especially for labeling Intervention and Outcome, the sentence-based approaches significantly outperform the BASELINE approach. Notice that aggregation via DS is only effective for the BASELINE annotations. This is expected since weighted aggregation methods rely on a certain noise level of the underlying annotations, which was high in the BASELINE (see Figure 6.3).

The results of aggregating all 8-17 annotations of the BASELINE approach are indicated by  $MV_{ALL}$  and  $DS_{ALL}$  in Table 6.4. As expected, the Kappa agreements substantially improve compared to the aggregation of only 3 annotations. However, for Intervention and Outcome, 8-17 aggregated annotations still reach substantially lower agreements to the gold standard than only 3 aggregated annotations of the SENSUPPORT approach, caused by the low quality of the underlying annotations (Figure 6.3). Only for Participant, the  $DS_{ALL}$  agreement of 0.867 significantly improves over all other aggregations of  $MV_3$  or  $DS_3$ . Regarding this result, we found that the DS algorithm picked up the signal from two workers of the BASELINE who annotated a majority of the abstracts with exceptionally high agreements to the gold standard of 0.83 and 0.84. These two workers' annotations were prioritized during aggregation, ignoring most of the other annotators.

### 6.4.3 Worker Feedback on the Usefulness of Dynamic Examples

We acquired feedback from the workers using the SENSUPPORT approach. We asked the workers at each task-instance if they found at least one of the dynamic examples useful. The feedback was obtained through an input form incorporated in the web-interface of the SENSUPPORT approach. Workers could select between the two labels *useful* and *not useful*, indicating whether the worker found at least one of the three presented dynamic examples useful or not.

We summarize the worker feedback on the usefulness of the dynamic examples in Table 6.5. The table shows that for most task-instances, the workers found at least one of the three dynamic

Table 6.4: Kappa agreements between aggregated annotations of each approach and the gold standard. We show significant improvements for both categories  $MV_3$  and  $DS_3$  where *a* refers to BASELINE and *b* to SENBASE (two-sided, paired t-test:  $p < 0.05$ ).

|                        | Cohen's Kappa ( $\kappa$ ) |                           |                           |
|------------------------|----------------------------|---------------------------|---------------------------|
|                        | P                          | I                         | O                         |
| BASELINE $_{MV_3}$     | 0.702                      | 0.455                     | 0.352                     |
| SENBASE $_{MV_3}$      | 0.715                      | 0.675 <sup>a</sup>        | 0.655 <sup>a</sup>        |
| SENSUPPORT $_{MV_3}$   | 0.780 <sup>ab</sup>        | <b>0.757<sup>ab</sup></b> | <b>0.694<sup>ab</sup></b> |
| BASELINE $_{DS_3}$     | 0.729                      | 0.579                     | 0.458                     |
| SENBASE $_{DS_3}$      | 0.726                      | 0.674 <sup>a</sup>        | 0.654 <sup>a</sup>        |
| SENSUPPORT $_{DS_3}$   | 0.776 <sup>a</sup>         | 0.756 <sup>ab</sup>       | <b>0.694<sup>ab</sup></b> |
| BASELINE $_{MV_{ALL}}$ | 0.760                      | 0.476                     | 0.343                     |
| BASELINE $_{DS_{ALL}}$ | <b>0.867</b>               | 0.633                     | 0.677                     |

examples useful, especially for the more difficult sub-task of labeling Intervention and Outcome. We further analyze whether the (perceived) usefulness of the examples affects the quality of the annotations. For that, we measure the Kappa agreement between the gold standard annotations and each worker’s annotated task-instances broken down for the usefulness of the dynamic examples. Table 6.6 shows that crowdworkers reach much higher agreements to the gold standard for task-instances where they found at least one of the dynamic examples useful than otherwise.

#### 6.4.4 Analysis of Agreement Types

We switch from analyzing Kappa agreements to analyzing which types of agreement appear between the gold standard and the non-expert annotators. The analysis of agreement types gives additional insights into the labeling behavior of annotators [LS19]. We differentiate between four agreement types, summarized in Table 6.7.

We first analyze how frequently workers entirely disagree with the gold standard in Figure 6.4a. Notice the substantial difference between the types Miss and Redundant when comparing the sentence-based approaches to the abstract-based approach BASELINE. In the BASELINE approach, we see a high frequency of Miss and fewer cases of Redundant for all PIO sub-tasks. On the other hand, in SENBASE and SENSUPPORT, we see a high frequency of Redundant cases and much fewer cases of Miss. This result shows that (i) crowdworkers who annotate entire abstracts frequently overlook text phrases that *should* be annotated, and (ii) crowdworkers who annotate sentences tend to label text phrases that *should not* be annotated.

We analyze how often workers exactly or at least partially agree with the gold standard annotations in Figure 6.4b. We find that crowdworkers using the SENSUPPORT approach have the highest frequency of token-level agreement to the gold standard, followed by SENBASE and BASELINE. Furthermore, the frequency of Exact cases is constantly higher in the sentence-based approaches compared to the BASELINE, especially for labeling Intervention and Outcome. Our results show that crowdworkers of the sentence-based approaches are more likely to fully agree with the gold standard than crowdworkers of the BASELINE approach.





Table 6.5: Percentage of task-instances where workers using the SENSUPPORT approach found at least one of the three shown dynamic examples useful.

| Feedback   | Percentage |     |     |
|------------|------------|-----|-----|
|            | P          | I   | O   |
| Useful     | 64%        | 78% | 76% |
| Not useful | 36%        | 22% | 24% |

Table 6.6: Kappa agreement between the gold standard and the crowdworkers annotations broken down for both feedback options. The agreements are averaged over the workers of the SENSUPPORT approach that annotated at least 5% of the 423 test sentences.

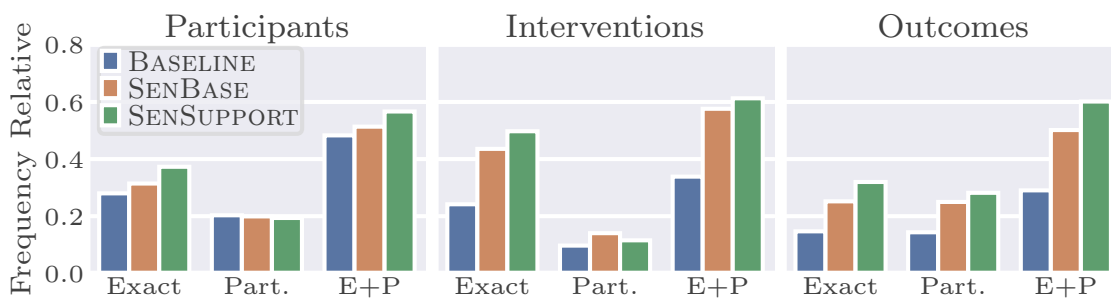
| Feedback   | Cohen’s Kappa ( $\kappa$ ) |          |          |
|------------|----------------------------|----------|----------|
|            | P                          | I        | O        |
| Useful     | 0.73±.12                   | 0.67±.14 | 0.60±.18 |
| Not useful | 0.42±.07                   | 0.41±.14 | 0.44±.18 |

Table 6.7: Overview of the differentiated agreement types. The examples show token-based text span annotations between crowdworkers (gray) and the gold standard (yellow).

| Type      | Example   |
|-----------|---|
| Exact     |  |
| Partial   |  |
| Miss      |  |
| Redundant |  |



(a) Frequency of crowdworkers not overlapping with the gold standard.



(b) Frequency of crowdworkers exactly or at least partially overlapping with the gold standard.

Figure 6.4: Relative frequency of the different agreement types between crowdworkers and the gold standard annotations. The combined result of Miss+Redundant and Exact+Partial is referred to as M+R and E+P, respectively.



## 6.5 Summary

We presented the DEXA annotation approach in which non-expert annotators are supported by dynamic examples that are semantically similar to the currently annotated sample. We evaluated the DEXA approach based on the PIO annotation task, where crowdworkers label the Participants, Interventions, and Outcomes in the text of clinical trial reports. We evaluate the following three annotation approaches:

- **SENSUPPORT**: This approach incorporates the DEXA approach by supporting workers with dynamic examples. The PIO annotations are assigned to sentences of clinical trial reports.
- **SENBASE**: In this approach, the PIO labels are annotated in sentences of clinical trial reports without the support of dynamic examples.
- **BASELINE**: The baseline approach of Nye et al. [NLP<sup>+</sup>18], where workers are asked to label PIO in entire abstracts of clinical trial reports.

We evaluated the sentence-based annotation approaches **SENBASE** and **SENSUPPORT**, and the abstract-based approach **BASELINE** by comparing crowd-annotations of each approach to a set of gold standard annotations. We found that crowdworkers using the **SENSUPPORT** approach reach the highest Kappa agreements to the gold standard. Therefore, whenever expert annotations can be spared, they should be utilized as dynamic examples. Furthermore, we showed that annotations from the sentence-based approaches **SENBASE/SENSUPPORT** reach substantially higher agreements to experts than annotations from the **BASELINE** approach, especially for the labels Intervention and Outcome. Therefore, we recommend splitting abstracts into sentences before being annotating in a crowdsourcing-based setting.

We further asked workers of the **SENSUPPORT** approach for explicit feedback on the usefulness of the dynamic example. For each annotated task-instance, we asked the crowdworkers whether one of the three presented dynamic examples was helpful in labeling the text. We found that workers perceive the proposed examples useful in 73% of the cases. For task-instances where workers found the dynamic examples useful, they reached on average a 0.24 higher Kappa agreement to experts (averaged over all PIO sub-tasks) than for samples where they did not find the dynamic examples useful.

Finally, we conducted a pairwise comparison of the token overlap of annotations of either approach with the gold standard. We found that crowdworkers using the sentence-based approaches are prone to annotate text phrases that should not be annotated, whereas workers using the abstract-based approach are prone to overlook text phrases that should be annotated. Overall, the highest frequency of token overlap agreement to the gold standard is reached by crowdworkers of the **SENSUPPORT** approach.

The collected annotations and the annotation interfaces of **SENBASE** and **SENSUPPORT** are available on GitHub<sup>6</sup>. The gold standard and **BASELINE** annotations can be found at<sup>7</sup>.

---

<sup>6</sup><https://github.com/Markus-Zlabinger/pico-annotation>

<sup>7</sup><https://github.com/bepnye/EBM-NLP>



# Conclusion

This thesis proposed novel approaches for improving the manual corpus annotation procedure. Our approaches support human workers in assigning annotations time-efficiently and with high accuracy, even for difficult, domain-specific tasks. In this chapter, we conclude the thesis by revisiting its contributions and research questions. Afterward, we discuss the limitations of our work and describe future research opportunities.

## 7.1 Research Questions and Contributions

We re-state the research questions asked in Chapter 1 and summarize the conducted research to answer the questions. Our annotation approaches incorporate unsupervised semantic short-text similarity (SSTS) methods. To identify and pick an effective method for our use-cases, we asked the following research question:

*How effective are (i) traditional, (ii) embedding-based, and (iii) contextualized unsupervised semantic short-text similarity methods for question-answering and biomedical sentence retrieval?*

In Chapter 4, we compared the effectiveness of ten unsupervised SSTS methods. We considered methods from three categories: traditional word-based similarity methods (e.g., TFIDF [BR99]), methods that aggregate word embeddings (e.g., Smooth Inverse Frequency [ALM17]), and methods that infer a contextualized text embedding (e.g., Sentence-BERT [RG19]).

We evaluated the methods on four benchmark datasets: Two biomedical sentence-to-sentence similarity datasets and two question-to-question similarity datasets. We applied the ten SSTS methods on the four benchmark datasets to measure their effectiveness in retrieving similar sentences and questions. Our results showed that the sentence embedding method Sent2Vec [PGJ18, CPL19] and the two weighted embedding-based methods (i.e., Weighted Averaging [LM14] and Smooth Inverse Frequency [ALM17]) are effective on all four benchmark datasets. We further found that these methods are independent of specific preprocessing, as they are equally effective whether the text is lowercased, lowercased with stopwords removed, or not preprocessed at all.

We used the evaluation of the various SSTS methods to identify effective methods needed to answer the other two research questions asked in this thesis. One of these questions addresses the time efficiency of the annotation procedure, defined as follows:

*How does pre-grouping of similar text samples impact the annotators' time efficiency for question-answering and biomedical named-entity recognition?*

In Chapter 5, we proposed the GROUP-WISE annotation approach, in which text samples (i.e., questions and sentences) are pre-grouped based on their semantic similarity before being labeled by human workers. We compared the efficiency of the GROUP-WISE annotation approach to the traditional approach of labeling each sample one by one, referred to as the SEQUENTIAL approach. We compared the two approaches for annotating customer-support questions and named-entities in biomedical publications. We grouped the text samples based on their similarity computed by an unsupervised SSTS method. We selected Sent2Vec [PGJ18, CPL19] for the biomedical annotation task and Smooth Inverse Frequency [ALM17] for the customer-support annotation task. To evaluate the annotation procedure, we monitored the annotators' average time and the number of user-interactions to label one sample. For both annotation tasks, we found that workers of the GROUP-WISE approach are quicker and require fewer user-interactions than workers of the SEQUENTIAL approach. Our results showed that pre-grouping similar text samples substantially improves the time efficiency of the conducted annotation tasks.

We further investigated whether the GROUP-WISE approach harms the label quality compared to the SEQUENTIAL approach. For that, we measured the label quality by computing the inter-annotator agreement between annotations of each approach and a set of gold standard annotations. We showed that annotations of both approaches reach similar agreements with the gold standard. We concluded that using the GROUP-WISE approach over the SEQUENTIAL approach does not harm the label quality.

The last research question of this thesis addresses the effectiveness of non-expert annotators for conducting difficult, domain-specific tasks. The question is defined as follows:

*How does showing examples similar to the currently annotated text sample—so-called dynamic examples—affect the label accuracy of non-expert annotators for biomedical named-entity recognition?*

In Chapter 6, we proposed the Dynamic EXamples for Annotation (DEXA) approach, in which we show examples to annotators that are semantically similar to the text currently annotated. We referred to the similar examples as *dynamic examples*, as they dynamically support annotators at each text sample. The aim of showing dynamic examples is to improve the label quality of non-expert annotators for difficult, domain-specific tasks.

We evaluated the effectiveness of the DEXA approach on the task of annotating the named-entities Participants, Interventions, and Outcomes in biomedical publications. The task was conducted by crowdworkers recruited from the Mechanical Turk crowdsourcing platform. We divided the crowdworkers into two groups: The first group was provided with task instructions and a few statically defined examples, and the second group was additionally supported by dynamic examples using the DEXA approach. The dynamic examples were retrieved from a small set of samples previously labeled by medical experts. We retrieved the similar examples using the semantic short-text similarity method Sent2Vec [PGJ18, CPL19], which we selected based on our comparative evaluation of the ten unsupervised SSTS methods. We measured the quality of each approach's annotations based on the inter-annotator agreement to a set of gold standard

annotations, assigned by medical experts. We found that crowdworkers supported by the DEXA approach reach significantly higher agreements to the experts than crowdworkers without such support. Our results demonstrated the effectiveness of dynamic examples to support non-experts during difficult, domain-specific annotation tasks.

## 7.2 Future Research

The proposed annotation approaches represent a foundation for future research projects aiming to improve the efficiency and the effectiveness of manual corpus annotation. Generalizing our approaches to new tasks, domains, and languages represents the opportunity to create new resources for systematic evaluation and supervised machine learning. For other researchers interested in re-using our work, we described our experiments in detail and made most of our data, code, and tools publicly accessible, with links available in each chapter's *Summary* section.

Another benefit of generalization is that our findings are further strengthened. This is important since a core challenge of experiments involving humans is reproducibility since humans are unique with different capabilities and qualifications to perform a certain task. The challenge of human uniqueness also reflects in manual corpus annotation, where two workers might assign annotations at a different pace or quality depending on their background. Therefore, it is crucial to test research hypotheses involving human annotation in the scope of large-scale experiments. We aimed to do so by evaluating the proposed annotation approaches over many annotators for different tasks.

Evaluating and comparing different annotation approaches for tasks critical to IR and NLP is another exciting research direction. Several annotation approaches are available, such as mentoring annotators, assigning annotation tasks based on the annotator's experience, or occasionally showing gold examples from which annotators learn how to perform the task. A comparative evaluation of annotation approaches is challenging and costly: We have to set up the experiment without inducing biases, and we need to recruit and compensate annotators suitable for the task. A cost-efficient evaluation approach is to assemble small-scale annotation experiments where only a small set of samples is annotated. Even from such small-scale experiments, scientific evidence can be obtained and used to derive best practices for future annotation projects.

This thesis showed that simple deviations from the typical manual corpus annotation procedure can lead to substantial improvements in effectiveness and time efficiency. We believe that our research only scratched the surface and that there are many other interesting research questions worth studying, such as the following: How are annotators trained effectively for a task? Should annotation instructions be concise or comprehensive? How should test runs be designed to identify accurate workers on crowdsourcing platforms? What annotation approaches or combination of approaches are most effective for a certain task? Answering these questions for IR and NLP-related tasks represents opportunities to contribute best practices and new approaches for improving manual corpus annotation.

### 7.2.1 Group-Wise Approach

We now discuss limitations and research opportunities specific to the GROUP-WISE annotation approach. This approach is suitable for tasks where annotators can label groups of similar samples more time-efficiently than labeling each sample individually. Tasks that potentially benefit from using the GROUP-WISE annotation approach are, e.g., the following:

- **Community Question Answering (CQA):** This task is a specific case of question-answering where users ask questions on public online boards, and the community then answers the questions. Well-known examples of CQA boards are websites like Quora<sup>1</sup> or Stack Overflow<sup>2</sup>. A common problem of CQA is that users tend to ask the same questions, making the community deal with questions that were already answered previously [HEABH17, CLG16]. The problem is addressed using duplicate detection systems, which are usually trained and evaluated on annotated corpora. These corpora can be created more time-efficiently by using the GROUP-WISE annotation approach.
- **Search Engines:** We require annotated datasets to measure the effectiveness of information retrieval systems. For creating new datasets for search engines, user queries and documents are manually assessed with respect to their relevance. The manual assessment's time efficiency can be improved by applying the GROUP-WISE annotation approach by pre-grouping queries with the same information need. Note that queries seeking the same information are characteristic of many search engines [WNZ02].
- **Autonomous Information Services:** These are systems where a user can autonomously request information. Such systems have emerged in the past decade to reduce the cost of traditional communication services, such as fax, e-mail, or telephone calls. Examples of autonomous information services are chat-bots for customer-support [Wan10], chat-bots for tourism information [PM16], or searchable FAQ databases [WYC05, KS08]. Users tend to repeatedly request the same information from autonomous information services, making the application of the GROUP-WISE annotation approach suitable.

The listed tasks represent future research projects for applying the GROUP-WISE approach. Note that the list of tasks is not exhaustive, as any task is eligible as long as similar samples appear, and the pre-grouping of these similar samples improves the time efficiency of annotators.

A viable research direction to further improve the GROUP-WISE approach is to develop new strategies to group similar samples. In our experiments, we used basic grouping strategies to keep the number of tunable parameters small, reducing the potential influence on our experiments' outcome. However, we believe that combining the GROUP-WISE approach with more sophisticated grouping strategies could further improve the approach's time efficiency. One such grouping strategy could be using clustering algorithms. Clustering algorithms take a set of vectorized texts as input and automatically generate clusters based on a predefined similarity function. The generated clusters can be used to derive groups of similar samples for the GROUP-WISE approach. However, be aware that using a clustering algorithm leads to new challenges, such as parameter tuning, cluster initialization, and measuring the quality of generated clusters.

### 7.2.2 Dexa Approach

A limitation of the DEXA approach is the availability of reference samples from which the dynamic examples are retrieved. Reference samples must be reliable with respect to correctness and are usually annotated by experts. Since experts are expensive and difficult to recruit, the number of reference samples from which dynamic examples are retrieved should be optimized: On the one hand, labeling too few reference samples can lead to them not covering frequent cases encountered during an annotation task. On the other hand, labeling too many reference samples can lead to redundantly labeled samples that do not increase coverage. Future research can address this

---

<sup>1</sup><https://www.quora.com/>

<sup>2</sup><https://www.stackoverflow.com/>

optimization problem by investigating the correlation between the number of reference samples and the annotation quality of non-experts supported by the DEXA approach.

Another research direction regarding the DEXA approach is in how reference samples are selected. In our experiments, we selected the reference samples randomly, fostering the chance of samples being redundant without contributing to coverage. However, we ideally want reference samples to cover as many cases as possible. The coverage could be improved by using a more sophisticated reference sample selection strategy. One direction could be combining the selection of reference samples with Active Learning [FZL13]. Active Learning is about prioritizing highly informative samples for manual labeling, which might also benefit the selection procedure of reference samples for the DEXA annotation approach.



Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.  
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

# Annotation Guidelines

This appendix contains the annotation guidelines used to prepare the annotators for the different tasks conducted in this thesis. The guidelines were used to train the annotators and align their conception on how the tasks should be performed. An overview of the different annotation guidelines used in the scope of this thesis is shown in Table A.1.

Table A.1: Overview of the annotation guidelines used in this thesis

| Task Description  | Used in Chapter | Guidelines |
|---|-----------------|------------|
| Relevance labeling between documents and search queries   | 2               | A.1        |
| Labeling the named-entities Participant, Intervention, and Comparison in clinical trial reports | 2               | A.2        |
| Assigning the polarity classes Positive, Neutral, and Negative to clinical study reports        | 2               | A.3        |
| Relevance labeling between diseases and symptoms  | 2               | A.4        |
| Labeling the named-entities Age, Gender, and Symptom in case study reports                      | 5               | A.5        |
| Labeling the named-entities Participant, Intervention, and Outcomes in clinical trial reports   | 6               | A.6        |

## A.1 Query-Document Relevance Labeling

### How to Annotate

Welcome to Fira! Our goal is to create fine-grained relevance annotations for query - passage pairs. In the annotation interface you will see 1 query and 1 passage and a range of relevance classes to select:



For each pair you must select 1 from 4 relevance classes:

- **Wrong** If the passage has nothing to do with the query, and does not help in any way to answer it
- **Topic** If the passage talks about the general area or topic of a query, might provide some background info, but ultimately does not answer it
- **Partial** The passage contains a partial answer, but you think that there should be more to it
- **Perfect** The passage contains a full answer: easy to understand and it directly answers the question in full

### Important Annotation Guidelines and Fira Usage Tips:

**(1)** You should use your general knowledge to deduce links between query and answers, but if you don't know what the question (or part of it such as an acronym) means, fall back to see if the passage clearly explains the question and answer and if not score it as **Wrong** or **Topic** only.

- We do not assume specific domain knowledge requirements.

- If the query makes no sense select **Wrong** (This might happen as we are dealing with real web queries).

**(2)** For **Partial** and **Perfect** grades you need to select the text spans, that answer the questions. You can select multiple words (the span) with your mouse or by once tapping or clicking on the start and once on the end of the span. You can select more than one span and you can also select them before clicking on the grade button. Below is an example of two selected spans:

#### difference between rn and bsn

The educational path for becoming a nurse vary depending on the type of nurse one hopes to become , but all nurses must be licensed . Nurse Types and Education Career Registered Nurse Licensed Practical and Licensed Vocational Nurses Educational Requirements Associate degree in nursing ( ADN ) , bachelor 's of science degree in nursing ( BSN ) or professional diploma from an approved nursing program Certificate from a 1 - year approved program Licensure Requirements Must pass the National Council Licensure Exam ( NCLEX - RN ) Must

The selection of words can be a bit tricky and multiple possible correct scenarios apply in many cases. To help you make better choices: Imagine that the words you select are extracted from the passage and displayed alone in a user interface or spoken by a digital assistant (without rewriting the answer: Extractive QA).

Therefore, select the span of words that sound most natural and compact (short) in answering the question.

For example:



- **Question asks for a date/location (when/where something happened)** If the passage contains the date/location and description, select the full date or the full location as answer.
- **Question asks for a definition** If the passage contains the definition only select the sentence(s) that a human would answer, without the boilerplate that many of these passages contain.
- **Question asks for a yes/no fact** If the passage does not contain yes/no as a word, but the fact stated, so that a human understands the meaning, select the fact span
- **Question asks for a number (with measurement)** Select the number and if followed by the measurement unit (meters, dollars ...) select also the unit word.
- **Question is answered by the whole passage** Select the whole passage (but try to do this as few times as possible).

(3) On the desktop you can use the keys 1-4 on your keyboard to quickly select the relevance label.

(4) You can see your annotation count and history to change an annotation (if you misclicked for example) via the dropdown menu in the upper-right corner.

Now before we get started, let's have a look at an example from each relevance grade:

#### causes of military suicide

Inside the Tortured Mind of Eddie Ray Routh, the Man Who Killed American Sniper Chris Kyle "In the Magazine U. S. Inside the Tortured Mind of Eddie Ray Routh, the Man Who Killed American Sniper Chris Kyle By Mike Spies On 11/23/15 at 12:22 PMChris Kyle, fourth from top left, was the most celebrated sniper in American military history. His killer, Eddie Ray Routh, may have been suffering from undiagnosed schizophrenia. Photo illustration: Joel Arbaje. Featured photos courtesy of Jodi Routh and AP. Share U. S. Chris Kyle U. S. Shootings Eddie Ray Routh This article first appeared on The Trace , an independent, nonprofit media organization dedicated to expanding coverage of guns in the United States.

1 Wrong

#### do goldfish grow

Caring for Your Goldfish in a Fish Bowl Without an Air Pump Pet Helpful » Fish & Aquariums » Freshwater Pets Caring for Your Goldfish in a Fish Bowl Without an Air Pump Updated on March 15, 2018Camile more Camile currently lives and works in the Middle East and has experience raising goldfish as a child. Contact Author Good aquarium plants are key to creating a healthy environment for goldfish when there isn't an air pump in the bowl. I currently live and work in the Middle East. One day, a friend gave me a goldfish in a bowl. At first, I was hesitant to accept the fish. I raised goldfish as a child, and I knew how much care they required.

2 Topic

**axon terminals or synaptic knob definition**



bodies are located in the ventral horn of the spinal cord. The terminal region of the axon gives rise to very fine processes that run along skeletal muscle cells. Along these processes are specialized structures known as synapses. The particular synapse made between a spinal motor neuron and skeletal muscle cell is called the motor endplate because of its specific structure.. "Figure 4.1 (see enlarged view)Consequently, an understanding of this synapse leads to an understanding of the others. Therefore, we will first discuss the process of synaptic transmission at the skeletal neuromuscular junction. The features of the synaptic junction at the neuromuscular junction are shown in the figure at left. Skeletal muscle fibers are innervated by motor neurons whose cell

3  
Partial

**causes of left ventricular hypertrophy**



Cardiovascular effects of hypertension Uncontrolled and prolonged elevation of BP can lead to a variety of changes in the myocardial structure, coronary vasculature, and conduction system of the heart. These changes in turn can lead to the development of left ventricular hypertrophy (LVH), coronary artery disease (CAD), various conduction system diseases, and systolic and diastolic dysfunction of the myocardium, complications that manifest clinically as angina or myocardial infarction , cardiac arrhythmias (especially atrial fibrillation), and congestive heart failure (CHF). Thus, hypertensive heart disease is a term applied generally to heart diseases, such as LVH (seen in the images below), coronary artery disease, cardiac arrhythmias, and CHF, that are caused by the direct or indirect effects of elevated BP.

4  
Perfect

## A.2 Biomedical Named-Entity Annotation

### Annotation of PICO

P<sub>opulation</sub> I<sub>ntervention</sub> C<sub>omparison</sub> O<sub>utcome</sub>

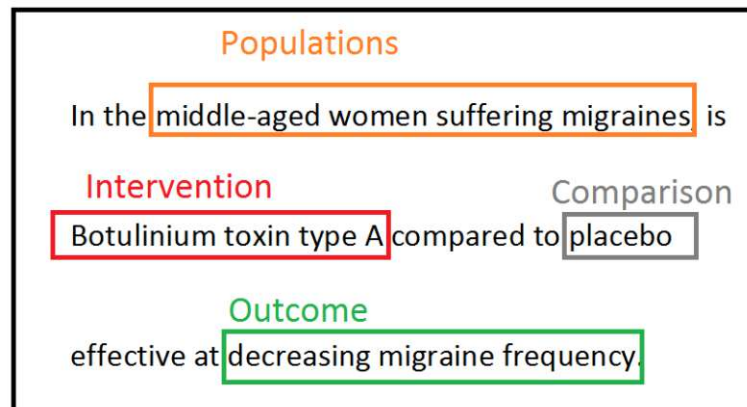
#### Introduction

For the KConnect project<sup>1</sup>, we seek to extract PICO elements in order to improve the medical search mechanism of PICO questions. The PICO questions aid medical searchers in order to rapidly find evidence based medical data (Table 1).

**Table 1:** Questions associated with each PICO element

| P  | I   | C  | O  |
|--|---|--|--|
| Patient, Population or Problem   | Intervention, observation or exposor                                      | Comparison   | Outcome  |
| What are the characteristics of the patient or population<br><br>What is the condition or disease you are interested in? | What do you want to do with this patient (e.g. treat, diagnose, observe)? | What is the alternative to the intervention (e.g. placebo, different drug, surgery)? | What are the relevant outcome (e.g. morbidity, death, complications) |

The PICO elements can be extracted and marked up on a phrase level in sentences (Figure 1).



**Figure 1:** Example of PICO element in a clinical question.

PICO is usually used as a Question and Answering (QA) system, where both the queries and the documents undergo Natural Language Understanding analysis, as seen in Table 2.

**Table 2:** QA example

|                             |   |
|-----------------------------|---|
| Question (Information Need) | I would like to know if it is ok to initiate a girl of 23 y old on Depo-Provera® who has been found to have vitamin D deficiency coincidentally. She does not seem to have any risk factors for vitamin D deficiency or osteoporosis. She is now on vitamin D |
|-----------------------------|---|

<sup>1</sup> <http://www.kconnect.eu/what-we-do>

| P                         | I            | C     | O                                    |
|---------------------------|--------------|-------|--------------------------------------|
| women vitamin d deficient | Depo-Provera | ----- | vitamin D deficiency or osteoporosis |

In order to model the QA solution, we have defined the task as a multi annotation model (Pustejovsky and Stubbs 2012). Each PICO element corresponds to a set of **semantic categories**, which we model according to (Huang et al 2006). The annotation model consists of five models, one for each PICO element and the sentiment of the outcome:

- The model **Population** is associated with a *people object*, such as humans and animals or part of humans and animals with an optional *patient qualifier* (e.g. gender, ethnic group and age). The People object is combined with a *physical condition* (e.g. healthy) and/or medical qualifier (e.g. medical history, diseases, symptoms, treatment status, treatments and drugs).
- The model **Intervention** is associated with treatments, drugs, procedures, diagnostic tests, exposures or observations and in some special cases symptoms (e.g. *a very low serum iron*).
- The model **Comparison**, is similar to the Intervention, i.e. comparing one intervention with a second intervention; however, the Comparison also includes placebo or non-treatment (Aspirin compared to no Aspirin) interventions.

Our annotation task requires more than one model to fully capture the information we need, which increases the complexity of the annotation task. Furthermore, the complexity increases due to the semantic categories, such as diseases, treatments and drugs. These categories can be part of different PICO elements, since the belonging depends on the syntactic context.

The semantic categories are in a sentence syntactically represented as noun phrases. These noun phrases (NP) are generally nested inside a preposition phrase (PP); and the PP is nested inside a parental noun phrase e.g. *patient with diabetes* [NP(patient(PP(IN with)NP(diabetes)))]]. Table 3 shows more complex examples of nested phrases for Population and Intervention.

**Table 3:** Example of different type of semantic categories for P and I

| Element      | Drug                     | Disease                               | Treatment                 | Example   |
|--------------|--------------------------|---------------------------------------|---------------------------|---|
| Intervention | Sorafenib                |                                       |                           | treatment <i>with</i> sorafenib   |
| Population   | ropivacaine/<br>fentanyl |                                       | artery bypass<br>grafting | patient-controlled epidural analgesia <i>with</i> ropivacaine/fentanyl <i>in off-pump</i> coronary artery bypass grafting |
| Population   |                          | metastatic<br>renal cell<br>carcinoma |                           | patients <i>with</i> metastatic renal cell carcinoma.   |

As seen in Table 3, the deciding component for labelling a semantic category as P or I depends on the context i.e. in different context a drug can be part of a population and in other as intervention. The PICO annotation task requires different level of linguistic information and also information from different medical domains, which further increase the complexity of the annotation task.

The PICO elements are associated with different aspects of noun phrases and domain terminology. The population elements generally consist of a noun phrase with one or more post modification e.g. *patient over forty with diabetes II*. Intervention and comparison consist of a multi word unit composed of general word and medical terms (e.g. maternal smoking). Outcome tend to occurs as noun phrase with an *of*-construction.

The domain terminology tend to be multi word units which consist of noun phrases containing common adjectives, nouns and occasionally prepositions (e.g. *of*). One of the major mechanisms of word formation is the morphological composite, which allows the formation of compound nouns out of two nouns (e.g. *floppy disk*, *blood cell*). The noun compounds could either be an orthographical unit (e.g. *bookcase*), or combined with hyphenation (e.g. *mother-in-law*) or a multi word unit (e.g. *crash landing*). Building blocks of the domain concepts are often common general English words, which are present in almost any text genre. But in a specific context and in a specific combination these noun phrases represent domain specific concepts. Medical terminology is a mixture of general words and morphemes related to Latin (Maglie 2009).

### Outline of the Guidelines

In the second phase of this annotation project we use two different group of annotators (medical librarians and computational linguistic students) in order to create a PIC corpus consisting of 2000 sentences. We have removed the Outcome for the annotation task due to the complexity of defining, identifying and interpreting what is the main outcome of a RCT or SR. The PIC corpus will be on phrase level, rather than purely on a sentence level and abstract level, as in Boudin *et al* (2010), Kim *et al* (2011), Wallace *et al* (2016).

To the second phase we choose to introduce an annotation confidence between [high confidence], [medium confidence] and [low confidence] for each annotated element. The *high confidence* is the default value and the last two requires that the annotator selects one them. We also limited the number for each PIC element to one per sentence.

In order to guide the annotators with only limited domain knowledge, we have labelled some terms with their respective semantic category, such as person, anatomy, disease and drug (Figure 2).

The screenshot shows a web-based annotation interface. At the top, there is a 'Title' field containing the text: 'Short term correction of anaemia with recombinant human erythropoietin and reduction of cardiac output in end stage renal failure.' Below this is an 'Abstract' section with a dropdown arrow. Navigation buttons include 'Previous Sentence' (with a dropdown showing '3'), 'Next Sentence', and 'Return to Overview'. The main area is titled 'Current Sentence (To annotate, click an arbitrary start and end token)' and displays a sentence with several terms highlighted and labeled with semantic categories: 'Person' (Eleven children), 'Disease' (end stage renal failure and anaemia), 'Disease' (haemoglobin concentration < 90 g/l), 'Anatomy' (cardiovascular), 'Disease' (anaemia), 'Person' (subcutaneous human recombinant erythropoietin), and 'Drug' (r-HuEpo). Below the sentence is an 'Annotations' table with three rows: 'Compare' (placebo), 'Population' (children with end stage renal failure and anaemia ( haemoglobin concentration < 90 g/l )), and 'Intervention' (subcutaneous human recombinant erythropoietin ( r-HuEpo )). Each row has a red trash icon on the right.

Figure 2: PIC annotation with pre-labelled semantic categories

As mentioned before, to conduct PIC labelling, an annotator needs to have medical domain knowledge. Additionally, linguistic knowledge is advantageous to understand the syntactic relations

between words. The syntactic relations in combination with the semantic categories reflect the PIC element for a specific context.

We are keeping the semantic labels, which are annotated with a GATE application. The GATE application does name entity recognition (NER), entity linking and disambiguation by using the KConnect knowledge base resource. However, the tool is not always correct as seen in Figure 3. The verb *aim* has been incorrectly identified as a noun belonging to the semantic category drug. The adjective *secondary* has been labelled incorrectly as disease and noun.

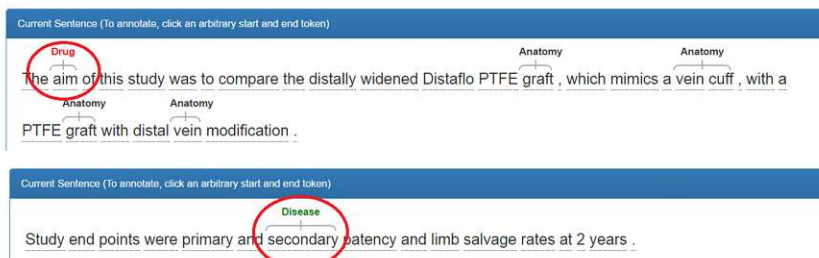


Figure 3: Examples of pre-label errors of semantic categories by the GATE application.

There is also cases where medical terms composed of MWU coincide with other medical terms e.g. *dosage arms* where *arms* has been label as anatomy.

From the first annotation round, we observed that annotators annotated differently due to their different background. The domain experts generally annotated the specific core element of the PIC elements more correctly than the linguists. Meanwhile, the linguists identified the entire nested noun phrase in which the core element occurs in. Both of described parts are relevant. For example, Population tend to be part of a prepositional phrase e.g. *in/for [PATIENT TYPE] with [MEDICAL CONDITION] between [AGE INFORMATION] undergoing [SOME SURGERY]*. Often only limited information is given about the Population (e.g. disease without a PEOPLE OBJECT), which makes the identification of the actual Population more difficult.

**Difficult to identify the entire scope of the population**

Comparing warfarin to aspirin (WoA) after *aortic valve replacement with the St. Jude Medical Epic heart valve bioprosthesis*: results of the WoA Epic pilot trial.

But by reading further in the abstract, you may find a sentence that captures the full scope of the Population.

“This prospective pilot study sought to investigate the feasibility of a larger trial and the efficacy of postoperative warfarin compared to acetyl salicylic acid (aspirin; ASA) **in patients after AVR with the St. Jude Epic porcine bioprosthesis (SJEP)**, and the feasibility of conducting a larger trial.”

Sometimes the indication semantic labels and people object can be ambiguous:

“A total of **40 fully veneered metal-ceramic crowns** were delivered in the posterior jaw[label:anatomy] segments of 20 patients using either a self-adhesive resin cement (RelyX

Unicem Aplicap, 3M ESPE; n = 20) or a zinc oxide phosphate cement (Hoffmann's Cement, Hoffmann; n = 20)”

In the above sentence we have the label anatomy for the word jaw, which is part of a noun phrase with an of construction i.e. *the posterior jaw segments of 20 patients*, but in this case the Population element is *40 fully veneered metal-ceramic crowns* since the focus of this clinical trial is on people with a dental prostheses.

Always make sure to include bracket information and sample size e.g. *in 249 patients with chronic kidney disease (CKD)* for Populations and dosage information for Interventions and Comparisons. But only if the information is directly connected to the PIC element.

### General Guidelines (relevant for all PIC elements)

#### Where is the PIC information usually located?

There are a few sentence types that occur frequently and contain much of the PIC information. These sentences are the following:

- Title sentence.
- In the first few sentences of the abstract:
  - Sentences that describe the study objective
    - *In this study, we want to compare Aspirin with placebo in patients with headache.*
    - *To evaluate the efficiency ...*
    - *The aim of this study ...*
    - *We aimed to assess ...*
  - Sentences that describe the study design
    - *239 patients with headache were randomized to either Aspirin or Placebo.*
    - *This blinded, randomized, ...*
    - *We assigned 239 patients with headache ...*
    - *We randomized ...*
    - *In this double-blind, ...*

#### Distinguish between sentences that refer to the current study rather and sentences that consist of general statements

You should not annotate PIC elements in general statement sentences (*"Aspirin in children with headache was covered in a previous study."*) but only in study referring sentences (*"In this study, we evaluate the effects of Aspirin in children with headache."*). Note that the first few sentences of an abstract may consist of general statement sentences (introducing the topic) and are then followed by a study referring sentence (describing the objectives or design of the current study).

#### Example1:



**General:** Children with end stage renal failure and anaemia have an increased cardiac index and often gross ventricular hypertrophy.

**Referring:** Eleven children with end stage renal failure and anaemia (haemoglobin concentration < 90 g/l) were enrolled into a single blind, placebo controlled, crossover study to assess ...

**Example2:**

**General:** Erlotinib, N-(3-ethynylphenyl)-6,7-bis(2-methoxyethoxy) quinazolin-4-amine is approved for the treatment for non-small cell lung cancer and pancreatic cancer.

**Referring:** Two phase I studies were conducted in healthy male subjects to evaluate the effect of pre- or co-administered rifampicin, a CYP3A4 inducer, on the pharmacokinetics of erlotinib.

Some keywords that indicate study referring sentences are the following:

"In this study...", "This study...", "We aimed...", "We assessed...", "We compared...", "To determine...", "To evaluate...", "We assigned...", "...were enrolled...", "...were randomized...", "...were assigned..."

POPULATION

**Sentence**

Only patients given NSAIDs continuously for at least 2 months with positive fecal occult blood (FOB) and endoscopically confirmed mild to moderate mucosal lesions (Lanza scale, grades 2-4) were included.

*Pattern*

[PATIENT TYPE] given [MEDICAL QUALIFIER] continuously for at least 2 months with [MEDICAL QUALIFIER] and [MEDICAL QUALIFIER] confirmed mild to moderate [MEDICAL QUALIFIER]

Annotate this word sequence as continuum for Population

patients given NSAIDs continuously for at least 2 months with positive fecal occult blood (FOB) and endoscopically confirmed mild to moderate mucosal lesions (Lanza scale, grades 2-4)

INTERVENTION & COMPARISON

Sentence: This double-blind, double-dummy, parallel-group study was designed to show that a pharmacokinetically enhanced formulation of oral amoxicillin-clavulanate (16:1, 2000/125 mg), twice daily, is at least as effective clinically and microbiologically as oral amoxicillin-clavulanate 1000/125 mg, three times daily, in the 10 day treatment of community-acquired pneumonia (CAP) in adults.

*Pattern*

a pharmacokinetically enhanced [INTAKE][DRUG][DOSAGE], [DOSAGE: duration] is at

least as effective clinically and microbiologically as [INTAKE][DRUG][DOSAGE],  
[DOSAGE: duration]

*Annotate this word sequence as continuum for Intervention*

a pharmacokinetically enhanced formulation of oral amoxicillin-clavulanate (16:1, 2000/125 mg), twice daily

*Annotate this word sequence as continuum for Comparison*

oral amoxicillin-clavulanate 1000/125 mg,three times daily

**Note!** The design of the trial is not part of the intervention or comparison element i.e. do not annotate double-blind, double-dummy, parallel-group

#### Guideline for the Model Population

The model **Population** is associated with a **patient type** i.e. animate entities such as human and animals or part of human and animals (e.g. anatomy). The patient type can be combined with different semantic categories, such as physical conditions (e.g. *healthy*), medical history (e.g. *with prior attacks of heart diseases*), diseases (e.g. *nonvalvular atrial fibrillation*), treatment status (e.g. *delayed treatment*), treatment and drugs (e.g. *taking hormone replacement therapy*), symptom (e.g. *chronic cough*) (Huang et al 2006). In addition to the more medical related categories the person object can be combined with a more detail description in term of a gender (e.g. *female*), age (e.g. *over forty*), national or ethnic belongings. Table 4 shows population examples, associated meta-categories and different combinations of semantic categories that represent the population element.

**Table 4:** Example of population and semantic categories

| POPULATION   | META-CATEGORY  | SEMENTIC CATEGORY                     |
|--|--|---------------------------------------|
| American Indian toddlers                               | patient qualifier + patient type                     | ethnic + age                          |
| hypertensives aged 55 or older                         | medical qualifier + patient qualifier                | symptom + age                         |
| children under 5 years with malaria                    | patient type + patient qualifier + medical qualifier | age + disease                         |
| healthy male volunteers                                | medical qualifier + patient qualifier + patient type | physical condition + gender           |
| pregnant women   | medical qualifier + patient qualifier                | physical condition + gender           |
| non-obese women with polycystic ovarian syndrome       | medical qualifier + patient type + patient qualifier | physical condition + gender + disease |
| patients on hemodialysis                               | patient type + medical qualifier                     | treatment status                      |
| renal transplant patients                              | medical qualifier+ patient type                      | treatment status                      |
| patients undergoing cesarean section                   | patient type + medical qualifier                     | treatment status                      |
| healthy humans   | medical qualifier + patient type                     | physical condition                    |
| HIV-uninfected adults                                  | medical qualifier + patient type                     | physical condition + age              |
| healthy volunteers                                     | medical qualifier + patient type                     | physical condition                    |
| Japanese patients with newly diagnosed Type 2 diabetes | patient qualifier + patient type + medical qualifier | ethnic + medical history + disease    |
| ocular hypertensive patients                           | medical qualifier + patient type                     | medical history                       |
| opioid-experienced volunteers                          | medical qualifier + patient type                     | medical history                       |

|   |  |                                  |
|---|--|----------------------------------|
| off-pump coronary artery bypass grafting                  | medical qualifier + (part of animate entities)       | Treatment                        |
| pterygium surgery   | medical qualifier                                    | Treatment                        |
| AAA surgery   | medical qualifier                                    | Treatment                        |
| patients with renal carcinoma                             | patient type + medical qualifier                     | Disease                          |
| hepatitis A   | medical qualifier                                    | Disease                          |
| chronic kidney disease                                    | medical qualifier                                    | Disease                          |
| employed depressed patients                               | patient qualifier + medical qualifier + patient type | Symptom                          |
| children between the age from 4 to 8                      | patient type + patient qualifier                     | category of human + specific Age |
| patient over 40 with diabetes                             | patient type + patient qualifier + medical qualifier | age + disease                    |
| women with cancer   | patient qualifier + medical qualifier                | gender + disease                 |
| American veterans   | Patient qualifier + patients type                    | Symptom                          |
| emergency department patients with chest pain and dyspnea | patient qualifier + patient type + medical qualifier | symptom + treatment status       |

On a meta-level, the element population is associated with three main categories (as seen in Table 4):

- patient type (e.g. *women, children*)
- patient qualifier (e.g. *American, aged 55 or older*)
- medical qualifier (e.g. *chronic kidney disease, AAA surgery, hypertensives, healthy pregnant, employed, HIV-uninfected*)

The three meta-categories can represent a population by itself or in combination with each other when they in the context answer to following question:

- *What are the characteristics of the patient or population?*

The population should be annotated as a continuum i.e. from the first identified word to the last word associated with the element population e.g. *patient with diabetes on insulin, woman with malaria and HIV positive*. Make sure to annotate all co-references associated with the first identified population e.g. the *Japanese patient with diabetes II* could later in the abstract be referred to *participant with diabetes II*. When you conduct the annotation, apply these rules according to the priority (1-3) rules:

1. If there is a **patient type** in the sentence combined with one or more **semantic categories** (see Table 4) annotate the full text phrase, which includes all meta-categories, as POPULATION.
2. If there is only a **patient type** in the sentence (e.g. *In this study only women participated*) without a patient qualifier or a medical qualifier, annotate only the patient type as POPULATION.
3. If there is only a **medical qualifier** in the sentence (e.g. disease: *in cancer*) without a **patient type or patient qualifier**, but with an intervention (e.g. *vitamin C in depression*), then annotate the medical qualifier as population and the intervention as intervention.

**However:** If you have identified a disease that is in relation to the population (e.g. "...patients with headache"), do not annotate the disease everywhere in the text as POPULATION it needs to be explicit or implicit related to the population element. For instance, let us assume we have the following two sentences in the abstract, we will get following annotations:

- **Sentence 1:** In this study, we observe the influence of Vitamin C in [*patients with headache*] POPULATION.
- **Sentence 2:** *Headache* plays a significant role in quality-of-life.

As seen in Sentence 1 there is an explicit relation, but the Sentence 2 contains only a medical condition but a weak link to population, therefore do not annotate it as population.

#### Guideline for the Model Intervention

The model **Intervention** is associated with different treatments: drugs (e.g. *warfarin*), procedures (e.g. *transvaginal ultrasound*), diagnostic tests (e.g. *pap smear*), symptoms (e.g. *a very low serum iron*) (Huang et al 2006). In less frequent cases, exposure or observation (e.g. *maternal smoking*) can be defined as Intervention. The Intervention answer the question:

- What do you want to do with this patient e.g. treat, observe, diagnose etc.

The Intervention is usually defined as the entity that the Population element is exposed to e.g. drug, observation, procedure or therapy. Table 5 shows different examples of interventions and their associated semantic categories.

**Table 5:** Example of intervention and semantic categories

| INTERVENTION   | SEMENTIC CATEGORY                             |
|--|---|
| Randomized phase II trial of first-line treatment with <b>sorafenib</b> versus interferon Alfa-2a in patients with metastatic renal cell carcinoma.  | Drug  |
| To evaluate the costs and benefits of potential <b>hepatitis A immunization</b> of healthy US children in regions with varying hepatitis A incidences.   | treatment (disease)                           |
| The effect of a <b>pain management program</b> on patients with cancer pain.   | Treatment                                     |
| The clinical benefit of <b>in-hospital observation</b> in 'low-risk' pneumonia patients after conversion from parenteral to oral antimicrobial therapy.  | Procedure                                     |
| Our aim was to assess the efficacy of <b>thoracic epidural anesthesia (EA) followed by postoperative epidural infusion (EI) and patient-controlled epidural analgesia (PCEA) with ropivacaine/fentanyl in off-pump coronary artery bypass grafting (OPCAB)</b> . | procedure (treatment + treatment + treatment) |
| Assessment of quality of life in patients on hemodialysis and the impact of <b>counseling</b> .  | Procedure                                     |
| Participants were randomly assigned to telephone interview only or to mail interview followed 2 weeks later by <b>telephone interview</b> .  | diagnostic test                               |
| We hypothesize that quantitative <b>pre-test probability</b> , linked to evidence-based management strategies, can   | diagnostic test                               |

|  |                 |
|--|-----------------|
| reduce unnecessary radiation exposure and cost in low-risk patients with symptoms suggestive of acute coronary syndrome and pulmonary embolism.  |                 |
| As part of a randomized control trial researching management of acute LRTi, an easy <b>self-completion diary</b> was formulated and validated against the 'measure yourself medical outcome profile 2' (MYMOP2), an instrument previously validated in general practice. | diagnostic test |

Identify the intervention in the title thereafter all co-reference to this particular element in the abstract. In the abstract it could be more detailed description see Figure 4.

Effects of pioglitazone in combination with metformin or a sulfonylurea compared to a fixed-dose combination of metformin and glibenclamide in patients with type 2 diabetes.

Abstract

Previous Sentence 2 Next Sentence Next Document

Current Sentence (To annotate, click an arbitrary start and end token)

Person Patients ( n = 250 ) treated with Drug metformin < or = 3 g/day ) or an SU as monotherapy for > 3 months and with glycosylated hemoglobin ( HbA ( 1c ) ) between 7.5 % and 11 % inclusive were randomized to receive either Drug pioglitazone ( 15-30 mg/day ) as add-on therapy to metformin or an SU or a fixed-dose combination of metformin ( Drug 400 mg ) and glibenclamide ( Drug 2.5 mg ) ( up to three tablets per day ) for 6 months .

Annotations

|              |  |  |
|--------------|--|--|
| Intervention | metformin  |  |
| Compare      | SU as monotherapy  |  |
| Compare      | glycosylated hemoglobin ( HbA ( 1c ) )                           |  |
| Intervention | pioglitazone   |  |
| Intervention | add-on therapy to metformin                                      |  |
| Compare      | SU   |  |
| Compare      | fixed-dose combination of metformin ( 400 mg ) and glibenclamide |  |

Figure 4: Example of co-reference occurrence of INTERVENTION element in title and abstract.

#### Guidelines: Model Comparison

The model **Comparison**, same as intervention but also including placebo, non-treatments and in rare cases can a disease be define as a comparison (e.g. *a flare-up of the Chron's*) (Huang et al 2006). The comparison element answer the question:

- What is the alternative to the intervention e.g. drugs, placebo, different drugs or surgery.

For example see Table 5-6 and Figure 5.

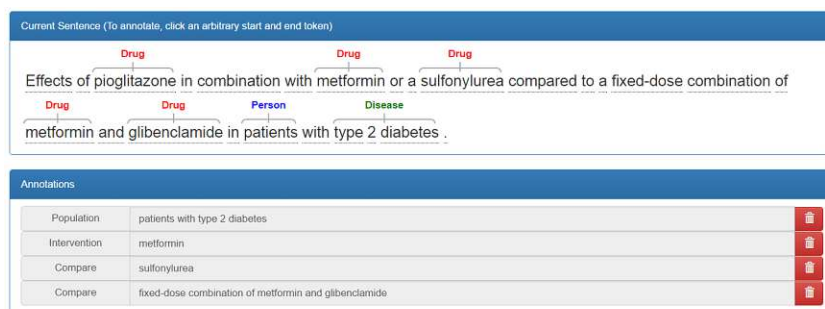


Figure 5: Example of C in relation to I element in a sentence.

Table 6: Example of comparison and semantic categories

| COMPARISON   | SEMENTIC CATEGORY |
|--|-------------------|
| Randomized phase II trial of first-line treatment with sorafenib versus <b>interferon Alfa-2a</b> in patients with metastatic renal cell carcinoma.  | drug              |
| Empagliflozin increased the rate and total amount of glucose excreted in urine compared to <b>placebo</b>  | drug              |
| <b>Everolimus</b> regimen compared with EC-MPS regimen is associated with lower incidence of DGF, slightly better 1-year graft survival rate, a significantly higher GFR and lower systolic blood pressure.  | drug              |
| Three AI/AN tribes were randomly assigned to two active interventions; a community-wide intervention alone (tribe A; n = 63 families) or <b>community-wide intervention containing a family component</b> (tribes B and C; n = 142 families).                              | procedure         |
| To compare the degree of conjunctival autograft inflammation, subconjunctival haemorrhage (SCH) and graft stability following the use of <b>sutures</b> or <b>fibrin glue (FG)</b> during pterygium surgery.   | treatment         |
| Participants (n = 160) were randomized to a 3-month group or <b>individual intervention utilizing a crossover design</b> .   | procedure         |
| Participants were randomly assigned to telephone interview only or to <b>mail interview</b> followed 2 weeks later by telephone interview.   | diagnostic test   |
| As part of a randomized control trial researching management of acute LRTi, an easy self-completion diary was formulated and validated against the ' <b>measure yourself medical outcome profile 2' (MYMOP2)</b> , an instrument previously validated in general practice. | diagnostic test   |

The comparison is usually defined in relation to the intervention using terms such as *or*, *versus*, *in comparison with* etc. The comparison is defined as an entity, which only a sub part of the population is exposed to. Identify the comparison by first identifying its relation to intervention in the title and abstract and thereafter all co-reference to this particular element in the text.

#### References

Boudin, F., Nie, J. Y., Bartlett, J. C., Grad, R., Pluye, P., & Dawes, M. (2010). Combining classifiers for robust PICO element detection. *BMC medical informatics and decision making*, 10(1), 29.

Huang X, Lin J, Demner-Fushman D. (2006) Evaluation of PICO as a knowledge representation for clinical questions AMIA Annu Symp Proc, 359-63.

Kim, N. S., Martinez D., Cavedon L., Yencken L. (2011) Automatic classification of sentences to support evidence based medicine. *BMC bioinformatics* 12.2: S5.

Maglie R. (2009) *Understanding the language of Medicine*, Aracne editrice

Pustejovsky, J. and Stubbs A. (2012). *Natural language annotation for machine learning*. O'Reilly Media, Inc.

Wallace, C. B., Kuipier J., Sharma A., Mingxi B. Z., Marshall I. . (2016) Extracting PICO sentences from clinical trial reports using supervised distant supervision. *Journal of Machine Learning Research* 17.132: 1-25.



## A.3 Clinical Study Polarity Analysis

Sentiment Annotation Guidelines

21.06.17

### Part1: Select the Conclusion Sentences

If the current abstract contains a conclusion section, skip the following instructions and go directly to "Part2: Annotate the Conclusion Sentences" of this guideline document.

If there is no conclusion section, you need to select where the conclusion starts. In this section, it is described how that is done.

#### How to select the conclusion?

The conclusion is located at the end of an abstract (usually the last 1-3 sentences). Therefore, in the annotation interface, select the first conclusion sentence of the abstract. All following subsequent sentences will be considered as a conclusion sentence as well.

#### Characteristics of the conclusion:

- Located at the end of an abstract
- Starts after the study results were described
- The conclusion is a summary of the described study and may contain following content:
  - the Population
  - the Intervention
  - the Comparison
  - the core results
  - For example: "Aspirin showed increased pain reduction effects than placebo in patients with headache."
- Can be a cocatenation of several sentences, e.g. if there are two core results.
  - For example: "Aspirin showed increased pain reduction effects than placebo in patients with headache. However, at the same time Aspirin resulted in negative side effects for a sub-group of patients."
- May contain specific keywords that indicate the start of the conclusion
  - "We conclude, ..."
  - "These results suggest ..."
  - "In conclusion, ..."

When selecting the beginning of the conclusion, the most difficult part is to distinguish between study results and conclusion. The study results may be very similar to the conclusion; however, are more specific and are usually based on some statistical evaluation results. In the conclusion, these study results are then summarized to a kind of main result of the given study.

Page 1 of 5

## Examples

To get an intuition on the boundary between conclusion and non-conclusion, consider the following examples. Note that these are snippets of full abstracts and the conclusion sentences are marked with yellow background.

*At 1 year and 2 years, the limb salvage rate was 72% and 65% for the precuffed group and 75% and 62% in the vein cuffed group ( $p = 0.88$ ). Although numbers are small and follow-up short, this midterm analysis shows similar results for the Distaflo precuffed grafts and PTFE grafts with vein cuff. A precuffed graft is a reasonable alternative conduit for infragenicular reconstruction in the absence of saphenous vein and provides favorable limb salvage.*

*Blood pressure did not change in six normotensive children completing an r-HuEpo limb; the decrease in cardiac index was therefore balanced by an increase in peripheral vascular resistance. Three children were taking anti-hypertensive treatment at the start of the study; one required an increase, and one a decrease, in treatment during the r-HuEpo limb. Short term treatment with r-HuEpo reduces cardiac index. A longer study is needed to determine whether this will, in time, result in a significant reduction in left ventricular hypertrophy.*

*The clinical, bacteriological and radiological success rates at the end of therapy (days 11-17) for the PP populations were all over 85%. Both regimens were well tolerated, with no differences in adverse events between the groups. Amoxicillin-clavulanate 2000/125 mg, twice daily, is well tolerated and at least as effective clinically as amoxicillin-clavulanate 1000/125 mg, three times daily, in patients with CAP and may also be appropriate for the treatment of infections due to *S. pneumoniae* strains with high-level penicillin resistance.*

*These alterations in breathing pattern were associated with CO<sub>2</sub> retention. Respiratory changes were mainly induced by the first injection of either drug. Despite increased plasma drug concentrations, subsequent doses did not cause further changes in respiratory variables except for an increase in PCO<sub>2</sub> after the second dose of midazolam. The clinical significance of these changes in PaCO<sub>2</sub> in otherwise healthy individuals is probably limited. The duration of the subjective sensation of sedation was longer after diazepam than after midazolam.*

*The differences between the treatments were particularly evident for men with minimal disease and good performance status; however, further studies should be conducted in this subgroup. Symptomatic improvement was greatest during the first 12 weeks of the combined androgen blockade, when leuprolide alone often produces a painful flare in the disease. We conclude that in patients with advanced prostate cancer, treatment with leuprolide and flutamide is superior to treatment with leuprolide alone.*

*Nitrous oxide concentration in the Reinforced, Sheridan, or Trachelon groups was slightly but significantly higher than that in the Profile or Hi-Contour groups. Cuff pressure never exceeded 22 mmHg and there were no air leaks. Therefore, inflating cuffs with 40% N<sub>2</sub>O preserves stable cuff pressure in all five tracheal tubes, despite differences in cuff and pilot balloon design.*

*Significant differences between the baseline examination and the follow-up examinations for sulcus bleeding index ( $p = 0.0013$ ) and plaque index ( $p < 0.0001$ ) were observed regardless of the luting agent used. The two cement types showed scarcely any differences between the parameters investigated. The outcomes of cementing fully veneered metal-ceramic crowns were equally good with self-adhesive resin cement as with the clinically proven zinc oxide phosphate cement.*

## Part2: Annotate the Conclusion Sentences

After selecting the conclusion starting point, each conclusion sentence should be displayed in the interface. Now, for each sentence, a sentiment is selected (Positive, Negative or Neutral).

When you annotate a sentence, keep attention to the wording, which may already indicates the sentiment of a sentence. For example:

### Positive sentiment phrasing:

- "...reasonable alternative..."
- "...significant improvement..."
- "...increase of Quality-of-Life..."

### Negative sentiment phrasing:

- "...increased mortality..."
- "...did not improve..."
- "...no significant improvement over placebo..."

### Positive Sentiment

In the following cases, a conclusion sentence should be annotated as Positive:

- Alternative or bioequivalence between Intervention and Comparison (If the goal of the study was to show that TreatmentA can be used as replacement for TreatmentB).
  - "A precuffed graft is a reasonable alternative conduit for infragenicular reconstruction in the absence of saphenous vein and provides favorable limb salvage."
- An Intervention is significantly effective.
  - "Our study showed that memantine is a tolerable and efficacious add-on treatment for primary negative symptoms of schizophrenia."
- An Intervention is significantly more effective than it's Comparison.
  - "Minocycline treatment for 3 months in children with FXS resulted in greater global improvement than placebo."
- The sentence contains a negative and a positive sentiment whereby the positive sentiment seems more important.
  - "This study confirms the possibility of obtaining an erectile response by intracavernous injection of 10 mg of moxisylyte with a very low incidence of local and systemic adverse effects."

### Negative Sentiment

In the following cases, a conclusion sentence should be annotated as Negative:

- An Intervention is not effective
  - *"Duloxetine 30 mg/d did not significantly reduce pain severity in patients with fibromyalgia."*
- An Intervention is less effective than it's Comparison. Note that if a treatment is equally effective as placebo or no-Intervention (i.e. drug taken vs non-taken), the sentence should be annotated as Negative.
  - *"In conclusion, treatment with ziprasidone monotherapy was not associated with any statistically significant advantage in efficacy over placebo."*
- The sentence contains a negative and a positive sentiment whereby the negative sentiment seems more important.
  - *"Although this trial did not report a benefit of inhalation aromatherapy for reducing anxiety, nausea, or pain when added to standard supportive care, it provides the first experimental rather than descriptive report on testing a single therapeutic essential oil among children and adolescents undergoing stem cell infusion."*

### Neutral Sentiment

In the following cases, a conclusion sentence should be annotated as Neutral:

- An Intervention is equally effective as it's Comparison.
  - *In this study, the confirmatory end point showed neutral results between the treatment groups.*
- The sentence contains a negative and a positive sentiment whereby both sentiments seem equally important.
  - *Performing coronary grafts on aspirin is associated with increased postoperative bleeding but may decrease the long-term hazard of coronary events.*
  - *Oxycodone was more potent than morphine for visceral pain relief but not for sedation.*
- All other sentences that can not be clearly assigned to either positive or negative

**Annotation Examples (green = positive, red = negative, gray = neutral)**

*Additional gait training may improve balance and gait performance and may induce changes in corticomotor excitability.*

*Short term treatment with r-HuEpo reduces cardiac index. A longer study is needed to determine whether this will, in time, result in a significant reduction in left ventricular hypertrophy.*

*In this target population selected according to positive FOB test and endoscopic evidence of mucosal injury, chronic administration of sucralfate significantly decreased NSAID-induced gastric erosions.*

*Short-term TNFalpha antagonism with infliximab did not improve and high doses (10 mg/kg) adversely affected the clinical condition of patients with moderate-to-severe chronic heart failure.*

*In a sample of pediatric ED patients with difficult access, ultrasound-guided intravenous cannulation required less overall time, fewer attempts, and fewer needle redirections than traditional approaches.*

*The early results of this WoA Epic pilot trial did not support the suggestion that patients receiving the SJEP, and tissue valves in general, should be administered warfarin to prevent valve thrombosis and peripheral arterial embolic phenomena.*

*In conclusion, both combined mediator blockade and combined topical corticosteroids are equally effective antiasthma therapy in patients with asthma and SAR.*

*The outcomes of cementing fully veneered metal-ceramic crowns were equally good with self-adhesive resin cement as with the clinically proven zinc oxide phosphate cement.*

*Bupropion was well tolerated and produced significantly greater-albeit quite modest-short-term weight loss in overweight and obese women with BED. Bupropion did not improve binge eating, food craving, or associated eating disorder features or depression relative to placebo. Our findings do not support bupropion as a stand-alone treatment for BED. The preliminary findings regarding short-term weight losses suggest the need for larger and longer-term trials to evaluate the potential utility of bupropion for enhancing outcomes of psychological interventions that have demonstrated effectiveness for BED but fail to produce weight loss.*

*In this study, the confirmatory end point showed neutral results between the treatment groups. However, a favorable outcome trend was seen in the severely affected patients with ischemic stroke treated with Cerebrolysin. This observation should be confirmed by a further clinical trial.*

*The clinical significance of these changes in PaCO<sub>2</sub> in otherwise healthy individuals is probably limited. The duration of the subjective sensation of sedation was longer after diazepam than after midazolam.*

## A.4 Disease-Symptom Relevance Assessment

### Projekt zur Erstellung von Evaluierungsdaten

**Beschreibung:**

In Zusammenarbeit mit einer Österreichischem Startup haben wir auf der TU-Wien ein neues System entwickelt welches zur Extrahierung von Krankheit-Symptom zusammenhängen verwendet werden kann. Dieses System nimmt als Eingabe eine Krankheit und liefert als Ausgabe ein Liste von Symptomen, sortiert nach deren Wichtigkeit. Zur Berechnung der Wichtigkeiten haben wir Zusammenhänge zwischen Krankheiten und Symptomen anhand von 1,5 Millionen medizinischen Volltext-Publikationen automatisch analysiert. Nun wollen wir die Effektivität unseres Systems evaluieren. Dafür benötigen wir passende Evaluierungsdaten (erstellt von medizinischen Experten).

**Aufgabe:**

Wir haben eine Liste bestehend aus 20 Krankheiten und den dazugehörigen Symptomen erstellt. Sie finden die Liste im nächsten Tabellen-Blatt namens „Daten“. Ihre Aufgabe ist nun für jede Krankheit die Leitsymptome per Checkbox zu markieren. Sie können alle verfügbaren Quellen (Textbücher, Google Suche) verwenden um die Leitsymptome zu recherchieren. Weiters können Sie Kollegen befrage, wobei Kollegen die ebenfalls an diesem Projekt beteiligt sind nicht befragt werden sollten.

**Aufwandsentschädigung:**

Nach der Evaluierung des Systems werden wir die Resultate im Rahmen eines wissenschaftlichen Konferenz-Papers publizieren. Es wird sich um eine Publikation im Bereich Computer Science (Unterbereich: Information Retrieval) handeln. Gerne biete ich allen die an der Erstellung der Evaluierungsdaten mitgearbeitet haben eine Rolle als Co-Autor bei diesem Paper an.

**Sonstiges:**

- Bei Fragen können Sie mich gerne und jederzeit per E-Mail Kontaktieren unter [markus.zlabinger@tuwien.ac.at](mailto:markus.zlabinger@tuwien.ac.at)
- Eine Krankheit kann *keine*, *eines* oder *mehrere* Leitsymptome haben.
- Falls Ihnen Fehler in unsereren derzeitigen Daten auffallen (zB ein fehlendes Symptom), bitte ich Sie diese zu Dokumentieren

**Beispiel:**

| Krankheit (dt.) | Krankheit (engl.) | Symptom (dt.)                           | Symptom (engl.)        | Ist Leitsymptom?                    |
|-----------------|-------------------|---|------------------------|-------------------------------------|
| Tennisarm       | Tennis Elbow      | Schmerzen im Arm                        | Arm Pain               | <input type="checkbox"/>            |
|                 |                   | Eingeschränkte Ellenbogen Beweglichkeit | Limited elbow movement | <input type="checkbox"/>            |
|                 |                   | Schwellung der Arme                     | Swelling of arm        | <input type="checkbox"/>            |
|                 |                   | Muskelschmerzen                         | Myalgia                | <input type="checkbox"/>            |
|                 |                   | Schmerzen im Ellenbogen                 | Pain in elbow          | <input checked="" type="checkbox"/> |
|                 |                   | Schwächegefühl , Hand                   | Weakness of hand       | <input type="checkbox"/>            |

## A.5 Labeling of Age, Gender, Symptom in Case Reports

### Instructions (Click to expand)

Highlight the text parts in the shown sentences that describe:

- The **Age** of the patient
  - "A **24-year-old** man ..."
  - **Do not** highlight age indicators like "teenager", "young", "senior", ...
- The **Gender** of the patient
  - "A 24-year-old **female** patient ..."
  - "The **boy** was ..."
  - "**He** experienced ..."
  - "**Her** blood levels ..."
- The **Symptoms** of the patient.
  - Make sure to include all characteristics of a symptom (for example: duration, time of onset, location)
  - "... experienced **2 weeks of headache** in addition ..."
  - "...measured **high fever** in ..."
  - "... he had **abdominal pain in the upper right quadrant** ."
  - **Do not** highlight separators between listings of symptoms
    - "... had **headache** and **39 degree fever at time of diagnosis** ."
    - "... symptoms were **headache** , **fever** , and **severe nausea** ."

How to highlight?

- Click a word in a sentence to highlight it.
- To highlight multiple words, click and hold the mouse while moving over all words that you wish to highlight.
- If none of the shown sentences contains information that should be highlighted, check the checkbox.

Examples

A **55-year-old woman** presented with **sudden-onset itchy lesions on her arms and legs for almost 20 days**.

Bronchoscopic examination showed an enlarged non-collapsible horseshoe-shaped trachea.

Atrial myxoma , the commonest primary cardiac neoplasm , presents with symptoms of **heart failure** , **embolic phenomena** or **constitutional upset** .

An **14-year-old boy** was evaluated for **chronic cough** and **right lower lobe (RLL) mass**.

The patient is currently being treated with 6 months of anticoagulation with rivaroxiban.

Questioning elicited an additional history of **sore throat** and **mild** , **dry cough** .

## A.6 Labeling of Participant, Intervention, and Outcome in Clinical Trial Reports

We collected annotations for Participant, Intervention, and Outcome in the scope of three self-contained sub-tasks. Therefore, we also provided guidelines specific to each sub-task. The shown guidelines were used to train annotators of the SENSUPPORT approach. We used almost the same guidelines for training the annotators of the SENBASE approach, with the only difference that the last bullet point describing the dynamic examples was omitted.

### A.6.1 Participant

**Task Instructions**

- In medical studies, the efficacy of medical treatments is evaluated within a group of study participants.
- **We present to you a sentence of a study report in which your task is to highlight the text that gives information about the participants of the study. You can highlight text in the sentence by clicking on a start and end word. If no information about the participants is mentioned, mark the corresponding checkbox.**
- Relevant information about participants include:
  - gender
  - medical conditions (e.g. diseases, upcoming surgery)
  - location ("patients in Taiwanese Hospitals")
  - how many people were in the study
- Do not highlight:
  - participant mentions without relevant information ("Patients were divided into two groups." versus "Patients with diabetes were divided into two groups.")
- To give additional context, we show the study report in that the sentence appears. The report might be helpful, e.g., to identify that an abbreviation AD stands for *Alzheimer Disease*.
- For the sentence that you should highlight, we show 3 similar sentences as examples that are already highlighted by a medical expert. **If one of these 3 examples is helpful to you in highlighting your sentence correctly, select the corresponding radio button.** However, the similar examples might not be always helpful: They might contain text that is not highlighted but should be; and second, the example might not be similar to the sentence that you highlight.

**Examples**

A study on the efficacy of recombinant human endostatin combined with apatinib mesylate in **patients with middle and advanced stage non-small cell lung cancer.**

To investigate the role of nicotinamide adenine dinucleotide phosphate 4 (NADPH4,NOX4) and transforming growth factor-beta (TGF- $\beta$ ) involve in pathogenesis of airway remodeling in **chronic obstructive pulmonary disease (COPD).**

**A total of 270 patients with MCI** were enrolled in a 24-week, multicenter, randomized, double-blind, placebo-controlled study.

**Participants with mild-to-moderate AD (Mini-Mental State Examination score of 13-26) were recruited from December 1999 to November 2000 using clinic populations, referrals from community physicians, and local advertising.**

A PPARA Polymorphism Influences the Cardiovascular Benefit of Fenofibrate in **Type 2 Diabetes:** Findings From ACCORD Lipid.



## A.6.2 Intervention

### Task Instructions

- In medical studies, the efficacy of medical treatments (called interventions) is evaluated within a group of study participants.
- **We present to you a sentence of a study report in which your task is to highlight the text that describes the intervention(s) of the study. You can highlight text in the sentence by clicking on a start and end word. If no information about the interventions is mentioned, mark the corresponding checkbox.**
- Interventions are:
  - a specific drug ("Aspirin")
  - surgery ("inguinal hernia repair")
  - talking therapy ("cognitive behavioural therapy")
  - or even a lifestyle modification (diet change or change of the toothpaste the patients use)
  - Many studies will have a group of patients who receive a "control" treatment, such as "placebo", no treatment at all, or the current standard practice ("usual care"). These should be also highlighted as interventions.
- Do not highlight:
  - details on how the interventions are given to the patients (e.g. "orally" or "intravenous")
  - information about dosages (e.g. "325 mg")
  - information about frequency or duration ("twice daily", "6 monthly sessions")
  - intervention mentions without relevant information ("The drug was administered" versus "The drug **Aspirin** was administered.")
- To give additional context, we show the study report in that the sentence appears. The report might be helpful, e.g., to identify that an abbreviation *RET* stands for *Regular Exercise Therapy*.
- For the sentence that you should highlight, we show 3 similar sentences as examples that are already highlighted by a medical expert. **If one of these 3 examples is helpful to you in highlighting your sentence correctly, select the corresponding radio button.** However, the similar examples might not be always helpful: They might contain text that is not highlighted but should be; and second, the example might not be similar to the sentence that you highlight.

### Examples

The patients received either **azithromycin** (600 mg/d for 3 days during week 1, then 600 mg/wk during weeks 2-12; n = 3879) or **placebo** (n = 3868).

**Zinc lozenges**, 10 mg, orally dissolved, 5 times a day (in grades 1-6) or 6 times a day (in grades 7-12).

**Antihypertensive therapy** was started immediately after randomization in the active treatment group, but only after termination of the double-blind trial in the control patients.

We report findings of a pilot RCT for a **parent training intervention** with a focus on the development of joint attention skills and joint action routines.

Treatment consisted of **nitrendipine** (10-40 mg/d), with the possible addition of **enalapril maleate** (5-20 mg/d), **hydrochlorothiazide** (12.5-25 mg/d), or **both add-on drugs**.

Seventy-two people residing in National Health Service (U.K.) care facilities who had clinically significant agitation in the context of severe dementia were randomly assigned to **aromatherapy with Melissa essential oil** (N = 36) or **placebo (sunflower oil)** (N = 36).

### A.6.3 Outcome

#### Task Instructions

- In medical studies, treatments are tested within a group of study participants. To determine if a new treatment works, various outcomes are measured in the people who take part in the study.
- **We present to you a sentence of a study report in which your task is to highlight the text that gives information about the outcomes of the study. You can highlight text in the sentence by clicking on a start and end word. If no information about the outcomes is mentioned, mark the corresponding checkbox.**
- Outcomes contain:
  - outcomes measured in patients ("blood pressure", "weight")
  - outcomes regarding the intervention ("effectiveness", "safety", "costs")
  - the score on a medical test or questionnaire ("Quality of Life Scales")
  - positive or negative events in the patient groups ("quit smoking", "death", "pain reduction")
  - adverse reactions ("Garlic lowers **blood pressure**")
- Do not highlight:
  - numbers or results (e.g. "**10 patients** quit smoking")
  - interpretations of outcomes (e.g. "**quality of life** improved among patients")
  - outcome mentions without relevant information ("Various outcomes were measured." versus "Various outcomes (**QoL, safety**) were measured.")
- To give additional context, we show the study report in that the sentence appears. The report might be helpful, e.g., to identify that an abbreviation *QoL* stands for *Quality of Life*.
- For the sentence that you should highlight, we show 3 similar sentences as examples that are already highlighted by a medical expert. **If one of these 3 examples is helpful to you in highlighting your sentence correctly, select the corresponding radio button.** However, the similar examples might not be always helpful: They might contain text that is not highlighted but should be; and second, the example might not be similar to the sentence that you highlight.

#### Examples

*Effects of 12 weeks' treatment with a proton pump inhibitor on **insulin secretion, glucose metabolism and markers of cardiovascular risk** in patients with type 2 diabetes*

*Secondary end points included change in **upper-limb measures** (including the **Fugl-Meyer Assessment-Upper Extremity**).*

*There were **no serious adverse device effects**.*

*The primary outcome measures were the **score on the Irritability subscale of the Aberrant Behavior Checklist and the rating on the Clinical Global Impressions - Improvement (CGI-I) scale** at eight weeks.*

*More **early-stage cancer** (stages I and II, 54 vs. 10, respectively;  $P < 0.001$ ) and **stage IIIa cancers** (15 vs. 3, respectively;  $P = 0.009$ ) were found in the screening group than in the control group.*

***Mortality, causes of death, and lung cancer findings** are reported to explore the effect of computed tomography (CT) screening.*

# List of Figures

|      |   |    |
|------|---|----|
| 2.1  | The four stages of an annotation project [SBDS14] . . . . .   | 12 |
| 2.2  | The Model-Annotate-Model-Annotate (MAMA) cycle proposed in [PS12a] . . . . .  | 23 |
| 2.3  | User interface of the tasks query-document relevance assessment and text span labeling. For text span labeling, annotators mark the characters within the document that answer the query’s information need (shown in A). For relevance assessment, annotators rate the document’s relevance by selecting one of the four classes (shown in B). . . . .   | 33 |
| 2.4  | User interface of the PIC annotation task with following components: (A) sentence navigation, (B) active sentence (yellow background), (C) active sentence split into tokens, and (D) selection of the PIC label. The PIC label is assigned through a pop-up window shown as soon as a token range is selected within the active sentence. . . . .  | 34 |
| 2.5  | User interface of the task polarity analysis of clinical studies. The full abstract (shown in A) is presented to annotators who label the polarity of individual sentences. The currently selected sentence, highlighted by a yellow background, can be labeled as positive, neutral, or negative (shown in B). . . . .   | 34 |
| 2.6  | Annotations of primary symptoms were collected by providing sheets of paper with disease-symptom pairs to the annotators. The figure shows disease-symptom pairs for the disease periodontitis. Notice that we included the German names for diseases and symptoms since the annotators are German native speakers. . . . .   | 35 |
| 2.7  | Cohen’s Kappa agreements between pairs of annotators for the five tasks. The line in each box indicates the median agreement and the dot the average agreement. . . . .   | 36 |
| 2.8  | Sample of the task of query-document relevance assessment. This sample was labeled by 44 annotators as <i>perfect</i> , 33 as <i>partial</i> , 1 as <i>topic</i> , and 1 as <i>wrong</i> . . . . .  | 37 |
| 2.9  | Sample of the task of query-document text span labeling with a heat-map indicating the tokens that were labeled by the 87 annotators . . . . .  | 38 |
| 2.10 | Sample of the task of biomedical named-entity annotation with a heat-map indicating the tokens that were labeled by the 5 annotators as Participants . . . . .  | 39 |
| 5.1  | Annotation interface of the GROUP-WISE annotation tool for question-answering. The illustration shows: (A) the candidate question, (B) the ranked list of similar questions, (C) the answer catalog, and (D) the input of the answer label. Note that the questions and answers appearing in the illustration are fictional since we cannot publish the actual data due to restrictions by the data provider. . . . . | 69 |
| 5.2  | The top-5 most frequently assigned answers by the two annotators. For conciseness, we aggregated the full answer text to a few keywords for this plot. . . . .  | 70 |
| 5.3  | Comparison of the efficiency between the GROUP-WISE approach and the SEQUENTIAL approach based on the customer-support dataset . . . . .  | 71 |

|     |   |    |
|-----|---|----|
| 5.4 | The annotation interface for the Mechanical Turk platform for labeling Age, Gender, and Symptom in case study reports . . . . .   | 74 |
| 6.1 | Annotation interfaces developed for the Mechanical Turk platform . . . . .  | 85 |
| 6.2 | Annotators can examine the entire abstract text through an expansible window in the sentence-based approaches. In the illustrated example, the currently annotated sentence is the title of the abstract, indicated by the blue border. . . . . | 85 |
| 6.3 | Kappa agreements between individual crowdworkers and the gold standard. . . . .   | 86 |
| 6.4 | Relative frequency of the different agreement types between crowdworkers and the gold standard annotations. The combined result of Miss+Redundant and Exact+Partial is referred to as M+R and E+P, respectively. . . . .                        | 89 |

# List of Tables

|     |   |     |
|-----|---|-----|
| 1.1 | Example of the medical polarity annotation task. The aim of this task is to assess whether the sentence has a positive or negative polarity. The presented sentence is from the abstract of [ORLS11]. . . . .   | 4   |
| 1.2 | Overview of the annotated datasets created in the scope of this thesis. The labeled text sample are, depending on the task, pairs, questions, sentences, or abstracts of biomedical publications. Multiple annotators labeled each sample to allow the analysis of cross-annotator behavior, such as the inter-annotator agreement. . . . .                 | 8   |
| 2.1 | Common types to annotate text data . . . . .  | 14  |
| 2.2 | Interpretation intervals for the Cohen’s Kappa score $\kappa$ . . . . .   | 26  |
| 2.3 | Overview of the characteristics of the five annotation projects . . . . .   | 28  |
| 2.4 | Sentences with annotated polarity classes of <i>positive</i> , <i>negative</i> , and <i>neutral</i> . . . . .   | 30  |
| 2.5 | Statistics of the five annotated datasets . . . . .   | 36  |
| 2.6 | Labeled sentences of the biomedical polarity analysis task. For each sentence, we report the number of labels assigned by the five expert annotators for Positive (Pos.), Neutral (Neu.), and Negative (Neg.). . . . .  | 40  |
| 2.7 | Depending on the context, treatment methods need to be labeled as Population or Intervention. . . . .   | 40  |
| 2.8 | Average Cohen’s Kappa agreement with standard deviations per disease for the task of primary symptom labeling of disease-symptom pairs . . . . .  | 41  |
| 3.1 | State-of-the-art results reported for the SemEval 2016 [NMM <sup>+</sup> 16] and 2017 [NHM <sup>+</sup> 17] similar question retrieval dataset. . . . .   | 49  |
| 3.2 | State-of-the-art results reported for the BIOSSES [SÖÖ17] and MedSTS [WAL <sup>+</sup> 18b] sentence similarity dataset. . . . .  | 50  |
| 4.1 | Overview of the evaluated pre-trained models . . . . .  | 58  |
| 4.2 | Pearson correlation coefficient for the task of biomedical sentence-to-sentence similarity. The Pearson correlation is computed between the ground truth labels and the similarity score of the unsupervised SSTS methods. For each corpus, we highlight the overall best result <b>bold</b> and the best result per category by <u>underline</u> . . . . . | 60  |
| 4.3 | Mean Average Precision (MAP) for the task of similar question retrieval. For each corpus, we highlight the overall best result <b>bold</b> and the best result per category by <u>underline</u> . . . . .   | 62  |
| 5.1 | Sentences from case study reports with annotations for <u>Age</u> , <u>Gender</u> , and <u>Symptom</u> . . . . .  | 73  |
|     |   | 125 |

|     |   |    |
|-----|---|----|
| 5.2 | Comparison of the time efficiency between SEQUENTIAL and GROUP-WISE approach for annotating case study reports. The times are recorded during annotating 450 sentences (= 150 HITS) by once using the SEQUENTIAL and once the GROUP-WISE approach. We rounded the hours and seconds up to the next integer value. . . . .                               | 75 |
| 5.3 | The number of clicks for crowdworkers switching to the labels Age, Gender, and Symptom. The presented numbers are recorded over annotating the 450 sentences (= 150 HITS) by once using the SEQUENTIAL and once the GROUP-WISE approach. Fewer clicks indicate a reduced overhead for switching between the labels. . . . .                             | 76 |
| 5.4 | Cohen's Kappa agreement between expert annotations and aggregated crowdworker annotations of both approaches. . . . .   | 77 |
| 6.1 | Analysis of the word counts of abstracts versus sentences. We measure the word count based on tokenized text, excluding punctuation. The data basis of this analysis is the EBM-NLP corpus, described in Section 6.3. . . . .   | 82 |
| 6.2 | Text samples with dynamic examples for <a href="#">Participants</a> , <a href="#">Interventions</a> , and <a href="#">Outcomes</a> . Note that only the labels for either P, I, or O (depending on the sub-task) within the dynamic examples are visible to workers. The labels shown in the text samples should be highlighted by the workers. . . . . | 83 |
| 6.3 | Overview of the compared annotation sets . . . . .  | 84 |
| 6.4 | Kappa agreements between aggregated annotations of each approach and the gold standard. We show significant improvements for both categories $MV_3$ and $DS_3$ where $a$ refers to BASELINE and $b$ to SENBASE (two-sided, paired t-test: $p < 0.05$ ). . . .   | 87 |
| 6.5 | Percentage of task-instances where workers using the SENSUPPORT approach found at least one of the three shown dynamic examples useful. . . . .   | 88 |
| 6.6 | Kappa agreement between the gold standard and the crowdworkers annotations broken down for both feedback options. The agreements are averaged over the workers of the SENSUPPORT approach that annotated at least 5% of the 423 test sentences. . . .   | 88 |
| 6.7 | Overview of the differentiated agreement types. The examples show token-based text span annotations between crowdworkers (gray) and the gold standard (yellow). . .   | 89 |
| A.1 | Overview of the annotation guidelines used in this thesis . . . . .   | 97 |

# Bibliography

- [ALM17] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations (ICLR)*, 2017.
- [ALP<sup>+</sup>16] Linda Andersson, Mihai Lupu, João Palotti, Allan Hanbury, and Andreas Rauber. When is the Time Ripe for Natural Language Processing for Patent Passage Retrieval? In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16*, pages 1453–1462, New York, NY, USA, October 2016. Association for Computing Machinery.
- [AMB<sup>+</sup>19] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78. Association for Computational Linguistics, 2019.
- [Aro01] A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *Proceedings of the American Medical Informatics Association Symposium*, pages 17–21, 2001.
- [BAC07] Niranjana Balasubramanian, James Allan, and W. Bruce Croft. A comparison of sentence retrieval techniques. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07*, pages 813–814, New York, NY, USA, July 2007. Association for Computing Machinery.
- [BAPM15] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.
- [BCC<sup>+</sup>16] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A Human Generated MACHine Reading COMprehension Dataset. *arXiv:1611.09268 [cs]*, November 2016.
- [BGJM17] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5(0):135–146, June 2017.
- [Bib93] Douglas Biber. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4):243–257, January 1993.

- [BKL09] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009.
- [BLC19] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3606–3611, 2019.
- [BPE<sup>+</sup>19] Jon Brasey, Christopher Price, Jonny Edwards, Markus Zlabinger, Alexandros Bampoulidis, and Allan Hanbury. Developing a fully automated evidence synthesis tool for identifying, assessing and collating the evidence. *BMJ Evidence-Based Medicine*, pages bmjebm–2018–111126, August 2019.
- [BR99] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., 1999.
- [CCKP16] Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. How to Train good Word Embeddings for Biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174, Berlin, Germany, August 2016. Association for Computational Linguistics.
- [CD17] Delphine Charlet and Geraldine Damnati. SimBow at SemEval-2017 Task 3: Soft-Cosine Semantic Similarity between Questions for Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 315–319, Vancouver, Canada, August 2017. Association for Computational Linguistics.
- [CDA<sup>+</sup>17] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics.
- [CKS<sup>+</sup>17] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- [CLG16] Pedro Chahuará, Thomas Lampert, and Pierre Gancarski. Retrieving and Ranking Similar Questions from Question-Answer Archives Using Topic Modelling and Topic Distribution Regression. *arXiv:1606.03783 [cs]*, June 2016.
- [CLX<sup>+</sup>17] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4470–4478, 2017.
- [CMB11] Hamish Cunningham, Diana Maynard, and Kalina Bontcheva. *Text Processing with GATE*. The University of Sheffield, 2011.
- [CMY<sup>+</sup>20] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. Overview of the TREC 2019 deep learning track. *arXiv:2003.07820 [cs]*, March 2020.



- [CPL19] Qingyu Chen, Yifan Peng, and Zhiyong Lu. BioSentVec: Creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE, 2019.
- [CTIB15] Justin Cheng, Jaime Teevan, Shamsi T. Iqbal, and Michael S. Bernstein. Break It Down: A Comparison of Macro- and Microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 4061–4064, New York, NY, USA, April 2015. Association for Computing Machinery.
- [CYK<sup>+</sup>18] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal Sentence Encoder. *arXiv:1803.11175 [cs]*, April 2018.
- [DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, 2018.
- [DD15] Kerstin Denecke and Yihan Deng. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine*, 64(1):17–27, May 2015.
- [DDC12a] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. Zen-crowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *21st International World Wide Web Conference (WWW2012)*, pages 469–478, 2012.
- [DDC12b] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *Proceedings of the First International Workshop on Crowdsourcing Web Search*, volume 842, pages 26–30, 2012.
- [DDS<sup>+</sup>09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [DKBH16] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. Toward a Learning Science for Complex Crowdsourcing Tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2623–2634, San Jose, California, USA, May 2016. Association for Computing Machinery.
- [DKC<sup>+</sup>18] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Computing Surveys (CSUR)*, 51(1):7:1–7:40, January 2018.
- [DPL<sup>+</sup>13] Radu Dragusin, Paula Petcu, Christina Lioma, Birger Larsen, Henrik L. Jørgensen, Ingemar J. Cox, Lars Kai Hansen, Peter Ingwersen, and Ole Winther. FindZebra: A search engine for rare diseases. *International Journal of Medical Informatics*, 82(6):528–538, June 2013.
- [DS79] Alexander Philip Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

- [dVGS<sup>+</sup>18] E. P. García del Valle, G. Lagunes García, L. Prieto Santamaría, M. Zanin, E. Menasalvas Ruiz, and A. Rodríguez González. Evaluating Wikipedia as a Source of Information for Disease Understanding. In *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pages 399–404, June 2018.
- [FDSMM17] Simone Filice, Giovanni Da San Martino, and Alessandro Moschitti. KeLP at SemEval-2017 Task 3: Learning Pairwise Patterns in Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 326–333, Vancouver, Canada, August 2017. Association for Computational Linguistics.
- [FE17] Mark A. Finlayson and Tomaž Erjavec. Overview of Annotation Creation: Processes and Tools. In *Handbook of Linguistic Annotation*, pages 167–191. Springer, 2017.
- [FKSR16] Marc Franco-Salvador, Sudipta Kar, Thamar Solorio, and Paolo Rosso. UH-PRHLT at SemEval-2016 Task 3: Combining Lexical and Semantic-based Features for Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 814–821, San Diego, California, June 2016. Association for Computational Linguistics.
- [FKTC13] Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, page 14. ACM, 2013.
- [FSL<sup>+</sup>18] Oluwaseyi Feyisetan, Elena Simperl, Markus Luczak-Roesch, Ramine Tinati, and Nigel Shadbolt. An extended study of content and crowdsourcing-related performance factors in named entity annotation. *Semantic Web*, 9(3):355–379, January 2018.
- [FYT14] Meng Fang, Jie Yin, and Dacheng Tao. Active Learning for Crowdsourcing Using Knowledge Transfer. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1809–1815, 2014.
- [FZL13] Yifan Fu, Xingquan Zhu, and Bin Li. A survey on instance selection for active learning. *Knowledge and Information Systems*, 35(2):249–283, 2013.
- [GCC12] Joseph K. Goodman, Cynthia E. Cryder, and Amar Cheema. Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, 26(3):213–224, 2012.
- [Goy17] Naman Goyal. LearningToQuestion at SemEval 2017 task 3: Ranking similar questions by learning to rank using rich features. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 310–314, 2017.
- [GSR18] Genevieve Gorrell, Xingyi Song, and Angus Roberts. Bio-YODIE: A Named Entity Linking System for Biomedical Text. *arXiv:1811.04860 [cs]*, November 2018.
- [GTC<sup>+</sup>20] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *arXiv:2007.15779 [cs]*, 2020.

- [HB12] Amaç Herdağdelen and Marco Baroni. Bootstrapping a game with a purpose for commonsense collection. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):1–24, 2012.
- [HEABH17] Amir Hazem, Basma El Amal Boussaha, and Nicolas Hernandez. MappSent: A Textual Mapping Approach for Question-to-Question Similarity. In *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, pages 291–300, November 2017.
- [Her09] William Hersh. *Information Retrieval: A Health and Biomedical Perspective*. Health Informatics. Springer-Verlag, New York, third edition, 2009.
- [HLA<sup>+</sup>21] Sebastian Hofstätter, Aldo Lipani, Sophia Althammer, Markus Zlabinger, and A. Hanbury. Mitigating the Position Bias of Transformer Models in Passage Re-Ranking. In *European Conference on Information Retrieval (ECIR)*, 2021.
- [HLD06] Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. Evaluation of PICO as a Knowledge Representation for Clinical Questions. *AMIA Annual Symposium Proceedings*, 2006:359–363, 2006.
- [HLZH20] Sebastian Hofstätter, Aldo Lipani, Markus Zlabinger, and Allan Hanbury. Learning to Re-Rank with Contextualized Stopwords. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM)*, pages 2057–2060, 2020.
- [HZH19] Sebastian Hofstätter, Markus Zlabinger, and Allan Hanbury. TU Wien@ TREC Deep Learning’19–Simple Contextualization for Re-ranking. In *Text REtrieval Conference (TREC)*, 2019.
- [HZH20a] Sebastian Hofstätter, Markus Zlabinger, and Allan Hanbury. Interpretable & Time-Budget-Constrained Contextualization for Re-Ranking. In *24th European Conference on Artificial Intelligence - ECAI 2020*, 2020.
- [HZH20b] Sebastian Hofstätter, Markus Zlabinger, and Allan Hanbury. Neural-IR-Explorer: A Content-Focused Tool to Explore Neural Re-ranking Results. In *European Conference on Information Retrieval (ECIR 2020)*, pages 459–464. Springer, 2020.
- [HZS<sup>+</sup>20] Sebastian Hofstätter, Markus Zlabinger, Mete Sertkan, Michael Schröder, and Allan Hanbury. Fine-Grained Relevance Annotations for Multi-Task Document Ranking and Question Answering. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM ’20*, pages 3031–3038, New York, NY, USA, October 2020. Association for Computing Machinery.
- [Ide17] Nancy Ide. Introduction: The Handbook of Linguistic Annotation. In *Handbook of Linguistic Annotation*, pages 1–18. Springer, 2017.
- [JK02] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, October 2002.
- [JN14] David Jurgens and Roberto Navigli. It’s All Fun and Games until Someone Annotates: Video Games with a Purpose for Linguistic Annotation. *Transactions of the Association for Computational Linguistics*, 2:449–464, 2014.

- [JPS<sup>+</sup>16] Alistair EW Johnson, Tom J. Pollard, Lu Shen, H. Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.
- [JR57] Firth John Rupert. A synopsis of linguistic theory. *Studies in Linguistic Analysis*, 1957.
- [JSPW17] Ayush Jain, Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. Understanding workers, developing effective tasks, and enhancing marketplace dynamics: A study of a large crowdsourcing marketplace. *Proceedings of the VLDB Endowment*, 10(7):829–840, March 2017.
- [Kek05] Jaana Kekäläinen. Binary and graded relevance in IR evaluations: Comparison of the effects on ranking of IR systems. *Information Processing and Management: an International Journal*, 41(5):1019–1033, September 2005.
- [KMCY11] Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. Automatic classification of sentences to support Evidence Based Medicine. *BMC Bioinformatics*, 12(2):S5, March 2011.
- [KS08] H. Kim and J. Seo. Cluster-Based FAQ Retrieval Using Latent Term Weights. *IEEE Intelligent Systems*, 23(2):58–65, March 2008.
- [KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems*, pages 1097–1105, 2012.
- [LB16] Jey Han Lau and Timothy Baldwin. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86, Berlin, Germany, August 2016. Association for Computational Linguistics.
- [LBRD<sup>+</sup>20] David La Barbera, Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias. In *European Conference on Information Retrieval*, pages 207–214, 2020.
- [Lee05] Geoffrey Leech. Adding Linguistic Annotation. In *Developing Linguistic Corpora: A Guide to Good Practice*, pages 1–16. Oxford: Oxbow Books, 2005.
- [LK77] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.
- [LM14] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML'14*, pages II–1188–II–1196, Beijing, China, June 2014.
- [LS19] Grace E. Lee and Aixin Sun. A Study on Agreement in PICO Span Annotations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'19*, pages 1149–1152, Paris, France, 2019. ACM Press.

- [LSB<sup>+</sup>16] Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher H. Lin, Xiao Ling, and Daniel S. Weld. Effective Crowd Annotation for Relation Extraction. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 897–906, San Diego, California, June 2016. Association for Computational Linguistics.
- [LSS11] Florian Laws, Christian Scheible, and Hinrich Schütze. Active Learning with Amazon Mechanical Turk. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1546–1556, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.
- [LYK<sup>+</sup>19] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, September 2019.
- [MBC14] Laure Martin, Delphine Battistelli, and Thierry Charnois. Symptom extraction issue. In *Proceedings of BioNLP 2014*, pages 107–111, Baltimore, Maryland, 2014. Association for Computational Linguistics.
- [MBS09] Meinard Müller, Andreas Baak, and Hans-Peter Seidel. Efficient and robust annotation of motion capture data. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 17–26, 2009.
- [McH12] Mary L. McHugh. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282, October 2012.
- [MCKP17] Chris Madge, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. Experiment-Driven Development of a GWAP for Marking Segments in Text. In *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play, CHI PLAY '17 Extended Abstracts*, pages 397–404, New York, NY, USA, October 2017. Association for Computing Machinery.
- [MMS93] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, June 1993.
- [MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [MSB<sup>+</sup>14] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.
- [MSC<sup>+</sup>13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.
- [MXT06a] Tony McEnery, Richard Xiao, and Yukio Tono. Corpus annotation. In *Corpus-Based Language Studies: An Advanced Resource Book*. Taylor & Francis, 2006.
- [MXT06b] Tony McEnery, Richard Xiao, and Yukio Tono. *Corpus-Based Language Studies: An Advanced Resource Book*. Taylor & Francis, 2006.



- [NFFZ17] J. Ni, H. Fei, W. Fan, and X. Zhang. Automated Medical Diagnosis by Ranking Clusters Across the Symptom-Disease Network. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1009–1014, November 2017.
- [NHM<sup>+</sup>17] Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. SemEval-2017 Task 3: Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 27–48, Vancouver, Canada, August 2017. Association for Computational Linguistics.
- [NIDL11] Aurélie Névéal, Rezarta Islamaj Doğan, and Zhiyong Lu. Semi-automatic semantic annotation of PubMed queries: A study on quality, efficiency, satisfaction. *Journal of Biomedical Informatics*, 44(2):310–318, April 2011.
- [NLP<sup>+</sup>18] Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 197–207, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [NM10] Graham Neubig and Shinsuke Mori. Word-based Partial Annotation for Efficient Corpus Construction. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).
- [NMM<sup>+</sup>16] Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, abed Alhakim Freihat, Jim Glass, and Bilal Randeree. SemEval-2016 Task 3: Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 525–545, San Diego, California, June 2016. Association for Computational Linguistics.
- [NZLH05] Yun Niu, Xiaodan Zhu, Jianhua Li, and Graeme Hirst. Analysis of Polarity Information in Medical Text. In *AMIA Annual Symposium Proceedings*, volume 2005, pages 570–574. American Medical Informatics Association, 2005.
- [ORLS11] G. Ormel, L. Romundstad, P. Lambert-Jensen, and A. Stubhaug. Dexamethasone has additive effect when combined with ondansetron and droperidol for treatment of established PONV. *Acta Anaesthesiologica Scandinavica*, 55(10):1196–1205, 2011.
- [PCK<sup>+</sup>13] Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, 3(1):3:1–3:44, April 2013.
- [PGJ18] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.
- [PM16] S. Pathak and N. Mishra. Context aware restricted tourism domain question answering system. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pages 534–539, October 2016.

- [PRLP18] Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. EMRQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2357–2368. Association for Computational Linguistics, 2018.
- [PS12a] James Pustejovsky and Amber Stubbs. Annotation and Adjudication. In *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*, pages 105–154. O’Reilly Media, Inc., 2012.
- [PS12b] James Pustejovsky and Amber Stubbs. Applying and Adopting Annotation Standards. In *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*, pages 87–104. O’Reilly Media, Inc., 2012.
- [PS12c] James Pustejovsky and Amber Stubbs. The Basics. In *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*, pages 1–32. O’Reilly Media, Inc., 2012.
- [PS12d] James Pustejovsky and Amber Stubbs. Building Your Model and Specification. In *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*, pages 67–86. O’Reilly Media, Inc., 2012.
- [PS12e] James Pustejovsky and Amber Stubbs. Defining Your Goal and Dataset. In *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*, pages 33–52. O’Reilly Media, Inc., 2012.
- [PS12f] James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. O’Reilly Media, Inc., 2012.
- [PSM14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
- [PVG<sup>+</sup>11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [PYL19] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy, August 2019. Association for Computational Linguistics.
- [RDS<sup>+</sup>15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, December 2015.
- [RG19] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3980–3990, Hong Kong, China, 2019. Association for Computational Linguistics.

- [RG20] Nils Reimers and Iryna Gurevych. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *arXiv:2004.09813 [cs]*, April 2020.
- [RKK<sup>+</sup>11] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. *Icwsm*, 11:17–21, 2011.
- [RS10] Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer, 2010.
- [RZLL16] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv:1606.05250 [cs]*, June 2016.
- [SBB<sup>+</sup>14] Adish Singla, Ilija Bogunovic, Gábor Bartók, Amin Karbasi, and Andreas Krause. Near-optimally teaching the crowd to classify. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML'14*, pages II–154–II–162, Beijing, China, June 2014. JMLR.org.
- [SBDS14] Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 859–866, 2014.
- [ŠGP20] Lea Škorić, Anton Glasnović, and Jelka Petrak. A publishing pandemic during the COVID-19 pandemic: How challenging can it become? *Croatian Medical Journal*, 61(2):79–81, April 2020.
- [She18] Kim Bartel Sheehan. Crowdsourcing research: Data collection with Amazon’s Mechanical Turk. *Communication Monographs*, 85(1):140–156, 2018.
- [SKZ<sup>+</sup>20] Marta Sabou, Klemens Käschnar, Markus Zlabinger, Stefan Biffl, and Dietmar Winkler. Verifying Extended Entity Relationship Diagrams with Open Tasks. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 132–140, 2020.
- [SLK<sup>+</sup>19] Setu Shah, Xiao Luo, Saravanan Kanakasabai, Ricardo Tuason, and Gregory Klopper. Neural networks for mining the associations between diseases and symptoms in clinical notes. *Health Information Science and Systems*, 7(1), December 2019.
- [SLZ<sup>+</sup>19] Ying Shen, Yaliang Li, Hai-Tao Zheng, Buzhou Tang, and Min Yang. Enhancing ontology-driven diagnostic reasoning with a symptom-dependency-aware Naïve Bayes classifier. *BMC Bioinformatics*, 20(1):330, December 2019.
- [SOJN08] Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics, 2008.
- [SÖÖ17] Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. BIOSSES: A semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58, July 2017.



- [SPT<sup>+</sup>12] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. BRAT: A web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, 2012.
- [SRG<sup>+</sup>96] D. L. Sackett, W. M. Rosenberg, J. A. Gray, R. B. Haynes, and W. S. Richardson. Evidence based medicine: What it is and what it isn't. *BMJ*, 312(7023):71–72, January 1996.
- [STL<sup>+</sup>18] M. S. Silberman, B. Tomlinson, R. LaPlante, J. Ross, L. Irani, and A. Zaldivar. Responsible research with crowds: Pay crowdworkers at least minimum wage. *Communications of the ACM*, 61(3):39–41, February 2018.
- [SUKG13] Christin Seifert, Eva Ulbrich, Roman Kern, and Michael Granitzer. Text Representation for Efficient Document Annotation. *Journal of Universal Computer Science*, 19:383–405, January 2013.
- [TS20] Noha S. Tawfik and Marco R. Spruit. Evaluating sentence representations for biomedical text: Methods and experimental results. *Journal of Biomedical Informatics*, 104:103396, April 2020.
- [VAD04] Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 319–326, 2004.
- [VH02] Ellen M. Voorhees and Donna Harman. Overview of TREC 2002. In *Text Retrieval Conference (TREC)*, 2002.
- [VH03] Ellen M. Voorhees and Donna Harman. Overview of TREC 2003. In *Text Retrieval Conference (TREC)*, pages 1–13, 2003.
- [VH12] Ellen M. Voorhees and William R. Hersh. Overview of the TREC 2012 Medical Records Track. In *Proceedings of The Twenty-First Text REtrieval Conference TREC*, 2012.
- [WAF<sup>+</sup>18] Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang D. Liu. MedSTS: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, January 2018.
- [Wal18a] Byron C. Wallace. Automating Biomedical Evidence Synthesis: Recent Work and Directions Forward. In *Proceedings of the 3rd Joint Workshop on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2018) Co-Located with the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018)*, pages 6–9, 2018.
- [WAL<sup>+</sup>18b] Yanshan Wang, Naveed Afzal, Sijia Liu, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Sunyang Fu, and Hongfang Liu. Overview of the BioCreative/OHNLP challenge 2018 task 2: Clinical semantic textual similarity. *Proceedings of the BioCreative/OHNLP Challenge*, 2018, 2018.
- [Wan10] D. S. Wang. A Domain-Specific Question Answering System Based on Ontology and Question Templates. In *2010 11th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, pages 151–156, June 2010.

- [WFS<sup>+</sup>20] Yanshan Wang, Sunyang Fu, Feichen Shen, Sam Henry, Ozlem Uzuner, and Hongfang Liu. The 2019 n2c2/OHNLTP Track on Clinical Semantic Textual Similarity: Overview. *JMIR Medical Informatics*, 8(11), November 2020.
- [WNZ02] Ji-Rong Wen, Jian-Yun Nie, and Hong-Jiang Zhang. Query Clustering Using User Logs. *ACM Transactions on Information Systems*, 20(1):23, 2002.
- [WYC05] Chung-Hsien Wu, Jui-Feng Yeh, and Ming-Jun Chen. Domain-Specific FAQ Retrieval Using Independent Aspects. *ACM Transactions on Asian Language Information Processing*, 4(1):17, 2005.
- [XSM<sup>+</sup>18] Eryu Xia, Wen Sun, Jing Mei, Enliang Xu, Ke Wang, and Yong Qin. Mining Disease-Symptom Relation from Massive Biomedical Literature and Its Application in Severe Disease Diagnosis. In *45 - AMIA 2018 Annual Symposium*, pages 1118–1126, 2018.
- [XY12] Fei Xia and Meliha Yetisgen-Yildiz. Clinical corpus annotation: Challenges and strategies. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTextM'2012) in Conjunction with the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey*, 2012.
- [YCA<sup>+</sup>19] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Multilingual Universal Sentence Encoder for Semantic Retrieval. *arXiv:1907.04307 [cs]*, July 2019.
- [YNC08] Rong Yan, Apostol Natsev, and Murray Campbell. A learning-based hybrid tagging and browsing approach for efficient manual image annotation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.
- [ZABH18] Markus Zlabinger, Linda Andersson, Jon Brassey, and Allan Hanbury. Extracting the Population, Intervention, Comparison and Sentiment from Randomized Controlled Trials. *Studies in Health Technology and Informatics*, pages 146–150, 2018.
- [ZAH<sup>+</sup>18] Markus Zlabinger, Linda Andersson, Allan Hanbury, Michael Andersson, Vanessa Quasnik, and Jon Brassey. Medical Entity Corpus with PICO elements and Sentiment Analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, 2018.
- [ZCY<sup>+</sup>19] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 6(1):1–9, May 2019.
- [ZH17] Markus Zlabinger and Allan Hanbury. Finding Duplicate Images in Biology Papers. In *Proceedings of the Symposium on Applied Computing, SAC '17*, pages 957–959, 2017.
- [ZHRH20] Markus Zlabinger, Sebastian Hofstätter, Navid Rekabsaz, and Allan Hanbury. DSR: A Collection for the Evaluation of Graded Disease-Symptom Relations. In *European Conference on Information Retrieval*, pages 433–440. Springer, 2020.
- [Zla19a] Markus Zlabinger. Efficient and Effective Text-Annotation Through Active Learning. In *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19*, pages 1456–1456, New York, NY, USA, 2019. ACM.

- [Zla19b] Markus Zlabinger. Improving the Annotation Efficiency and Effectiveness in the Text Domain. In *European Conference on Information Retrieval (ECIR 2020)*, Lecture Notes in Computer Science, pages 343–347. Springer International Publishing, 2019.
- [ZMBS14] XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. Human symptoms–disease network. *Nature Communications*, 5(1), December 2014.
- [ZRZH19] Markus Zlabinger, Navid Rekabsaz, Stefan Zlabinger, and Allan Hanbury. Efficient Answer-Annotation for Frequent Questions. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, Lecture Notes in Computer Science, pages 126–137. Springer International Publishing, 2019.
- [ZSH<sup>+</sup>20] Markus Zlabinger, Marta Sabou, Sebastian Hofstätter, Mete Sertkan, and Allan Hanbury. DEXA: Supporting Non-Expert Annotators with Dynamic Examples from Experts. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20*, pages 2109–2112. Association for Computing Machinery, July 2020.
- [ZSHH20] Markus Zlabinger, Marta Sabou, Sebastian Hofstätter, and Allan Hanbury. Effective Crowd-Annotation of Participants, Interventions, and Outcomes in the Text of Clinical Trial Reports. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3064–3074, 2020.
- [ZW18] Minghua Zhang and Yunfang Wu. An Unsupervised Model with Attention Autoencoders for Question Retrieval. In *Thirty-Second AAAI Conference on Artificial Intelligence*, page 9, 2018.
- [ZWTM10] Jingbo Zhu, Huizhen Wang, Benjamin K. Tsou, and Matthew Y. Ma. Active Learning With Sampling by Uncertainty and Density for Data Annotations. *IEEE Trans. Audio, Speech & Language Processing*, 18(6):1323–1331, 2010.