TU WIEN Informatics

Rule Learning for Open Information Extraction

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Data Science

eingereicht von

Marina Sommer, BSc
Matrikelnummer 11778902

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Ass. Gábor Recski, PhD

Wien, 28. April 2025

Marina Sommer	Gábor Recski

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

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

TU WIEN Informatics

Rule Learning for Open Information Extraction

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Data Science

by

Marina Sommer, BSc
Registration Number 11778902

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Ass. Gábor Recski, PhD

Vienna, April 28, 2025

Marina Sommer	Gábor Recski

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

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

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

Erklärung zur Verfassung der Arbeit

Marina Sommer, 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.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT- Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 28. April 2025

Marina Sommer

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

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

Acknowledgements

First of all, I want to thank my supervisor, Gábor Recski, for his great expertise and support while working on my thesis. Thank you for always making time for my questions and for being so flexible with online meetings while I was traveling the world.

Special thanks to Telmo Menezes and Camille Roth for helping me with fundamental questions about `graphbrain` and for providing the unpublished script for your open information extraction. This was incredibly helpful for understanding your extraction method and adapting it to my needs.

I am deeply grateful to everyone who supported me during my studies, for listening to my complaints and for motivating me. I especially want to thank my parents, my brother, and Fabian - I'm so thankful to have you in my life!

Last but not least, I want to thank you, Nazish, for your motivating words when we met in Singapore during my travels. Your career is truly inspiring, and I admire your courage to follow your dreams!

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

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

Kurzfassung

Open Information Extraction (OIE) ist ein Teilbereich der natürlichen Sprachverarbeitung, bei dem Text automatisch in Tupel strukturiert wird, die Beziehungen zwischen einem Prädikat und mehreren Argumenten abbilden. Dies ist besonders nützlich für weiterführende Anwendungen, wie Frage-Antwort-Systeme, Textzusammenfassungen oder den Aufbau von Wissensdatenbanken. Obwohl neuronale Modelle derzeit die Forschung auf diesem Gebiet dominieren, schränkt deren Black-Box-Natur die Interpretierbarkeit ein und wirft Bedenken hinsichtlich Datenschutz und Datenverzerrungen auf. Darüber hinaus sind die enormen Rechenanforderungen dieser Modelle mit einem hohen Energieverbrauch und CO_2-Emissionen verbunden, was erhebliche Herausforderungen für die Nachhaltigkeit darstellt.

Diese Diplomarbeit präsentiert ein vollständig transparentes, regelbasiertes OIE-System, das auf dem Konzept semantischer Hypergraphen nach Menezes und Roth [MR21] aufbaut. Durch die Integration von Annotationen aus dem `LSOIE`-Datensatz wird deren Entwicklung erweitert und in ein "supervised rule learning" System überführt. Zusätzlich erlaubt die entwickelte Lösung eine flexible Anpassung verschiedener Parameter, wie etwa der Anzahl verwendeter Regeln für die Tupel Extraktion, und unterstützt eine Steuerung des Verhältnisses zwischen Precision und Recall, je nach Zielsetzung des Nutzers.

Das System erzielte eine wettbewerbsfähige Leistung bei der Evaluierung auf vier Testdatensätzen aus zwei unterschiedlichen Domänen. So wurde beispielsweise auf einer reduzierten Version des `LSOIE-sci/test`-Datensatzes eine Precision von 40% und ein Recall von 31,9% erreicht, was einem F_1-Wert von 0,355 entspricht. Die Ergebnisse zeigen, dass regelbasierte Methoden konsistente und interpretierbare Lösungen für OIE liefern können und somit eine praktikable Alternative zu komplexen und undurchsichtigen neuronalen Netzwerken darstellen. Die Implementierung dieser Arbeit ist als Open-Source in einem Fork von `newpotato` auf GitHub[1] verfügbar und wurde unter der MIT-Lizenz veröffentlicht.

[1] `https://github.com/whoopi24/newpotato` (Zuletzt zugegriffen: 17. April 2025)

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

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

Abstract

Open information extraction (OIE) is a natural language processing (NLP) task that automatically structures text into tuples, representing relations between a predicate phrase and several arguments. This is particularly useful for various downstream applications such as question-answering systems, text summarization and knowledge base construction. Although neural models currently dominate research in the field, their black box nature limits interpretability and raises concerns about data privacy and data biases. Additionally, the enormous computational demands of these models result in high energy consumption and carbon emissions, posing significant sustainability challenges.

This diploma thesis presents a fully transparent, rule-based OIE system that builds on Menezes and Roth's framework using semantic hypergraphs [MR21]. In particular, it extends their approach by incorporating annotations from the `LSOIE` dataset, transforming it into a supervised rule learning system. Furthermore, the resulting solution offers flexible parameters, such as the number of symbolic patterns used for tuple extraction, supporting a trade-off between precision and recall depending on the user's objective.

The system demonstrated competitive performance in evaluations across four test datasets from two distinct domains. For instance, for a filtered version of the `LSOIE-sci/test` dataset, it achieved 40% precision and 31.9% recall, leading to an F_1 score of 0.355. The overall results highlight that rule-based approaches can provide consistent and interpretable solutions for OIE, offering a viable alternative to complex and opaque neural networks. The implementation of this work is available open-source in a fork of the `newpotato` repository on GitHub[2], and is released under the MIT License.

[2] `https://github.com/whoopi24/newpotato` (Last accessed: April 17, 2025)

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

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

Contents

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

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

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

CHAPTER 1

Introduction

Open information extraction (OIE) is the task of mapping sentences to tuples (often termed triplets), which are ordered sets of elements representing relationships between entities. For example, the sentence "the cat chased the mouse" can be mapped to the triplet (the cat; chased; the mouse). Similarly, the sentence "the dog stole a sock from the laundry basket" can be mapped to (the dog; stole; a sock). This thesis explores how such tuples can be extracted using simple rules (also referred to as symbolic patterns) that are interpretable and explainable. For instance, a basic rule that works for the examples above might be:

```
If a predicate p connects two noun phrases n1 and n2,
the triplet (n1; p; n2) will be extracted.
```

Essentially, rules are the operational logic behind how OIE systems know *what* to extract from *where* in a sentence. In this context, rule learning refers to the process of automatically discovering and generating symbolic patterns from a text corpus, which are then used for tuple extraction. This procedure can be applied in either a supervised or unsupervised manner. One benefit of unsupervised learning is that no labeled data is required, as the derived patterns are purely based on the structure of words and phrases. In contrast, a supervised approach can be automated, eliminating manual effort.

This work builds on the OIE system proposed by Menezes and Roth [MR21], which is based on semantic hypergraphs (SHs). These graphs offer an effective representation of natural language, as explained in Section 2.2. Consider the sentence "Webster is described as Benefield's boyfriend", depicted in Figure 1.1 using SH notation, where each token is annotated with a specific type (e.g., predicate `/P`, concept `/C`, modifier `/M`, etc.).

In this regard, OIE patterns must conform to the SH structure. An example pattern in SH notation is `(REL/P ARG0/C ARG1/S)`, where:

Sentence: Webster is described as Benefield's boyfriend.

Semantic Hyperedge:

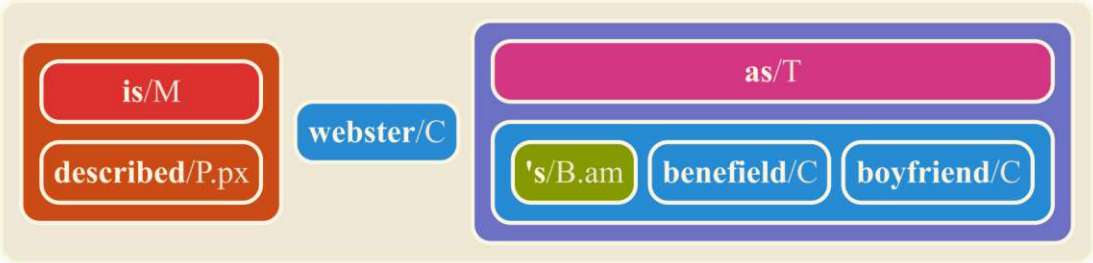


Figure 1.1: Example of Semantic Hyperedge representation.

- `REL/P` refers to the predicate or relation (e.g., "is described"),
- `ARG0/C` represents the subject (e.g., "Webster"), and
- `ARG1/S` refers to the object or complement (e.g., "as Benefield's boyfriend").

Ultimately, mapping the pattern `(REL/P ARG0/C ARG1/S)` to the semantic hyperedge shown in Figure 1.1 results in the extraction of the triplet (Webster; is described; as Benefield's boyfriend).

Modern NLP applications are increasingly driven by neural models, specifically large language models (LLMs). A major drawback of these systems is their black box nature, which makes the decision-making processes opaque and hard to interpret. This lack of transparency likely stems from their complexity introduced by a vast number of parameters. Bender et al. [BGMMS21] refer to LLMs as "stochastic parrots", since they generate text in a predictive manner without truly understanding its meaning. This behavior is akin to "parroting" patterns learned from their training data, raising concerns about inherent data biases. Additionally, both the training and deployment of these models demand immense computational resources, prompting questions about sustainability and accessibility. Consequently, NLP researchers are advised to thoroughly weigh the risks and costs of large language models and to develop effective NLP methods that are not excessively data-intensive. As outlined by H. Brown et al. [BLM+22], language models also entail data privacy risks, particularly when applied in sensitive sectors such as healthcare and finance, which users of LLMs should carefully consider.

Due to the aforementioned issues and risks of black box models, including unreliability and untrustworthiness, the aim of this thesis is to provide symbolic rules for the open information extraction task and to present strategies for building a rule-based system that is comparable to state-of-the-art approaches. The final OIE system is fully transparent and easy to understand, indicating that it offers complete visibility into how it processes

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

text and how its results are obtained. By always applying the same explicit rules, consistent performance can be expected. These characteristics should enhance trust into the system and facilitate user interaction and improvement. This thesis addresses the following three research questions:

1. **What strategies can be used to find accurate patterns for OIE? (RQ1)**
   This question explores the different techniques and approaches that can be employed to discover and develop accurate (effective) patterns for extracting information from unstructured text in the context of OIE. Effective patterns achieve high accuracy and can be reliably used across different datasets and scenarios. Ultimately, the focus lies on identifying strategies that enhance precision (how accurate the extractions are) and recall (how complete the extractions are).
2. **How do rule-based OIE solutions perform in comparison to state-of-the-art methods regarding precision and recall? (RQ2)**
   This question examines the performance of rule-based OIE systems in extracting information compared to the latest, most advanced methods in the field, with a focus on the two key metrics, precision and recall. This comparison aims to evaluate whether rule-based approaches can match or exceed the performance of modern techniques, or if there are specific strengths and weaknesses associated with each approach.
3. **How well do symbolic OIE patterns generalize across various datasets compared to state-of-the-art methods? (RQ3)**
   This question investigates how well the extraction patterns used by rule-based OIE systems maintain their performance when applied to a diverse range of datasets. It concentrates on evaluating the ability of symbolic patterns to generalize effectively across different types of texts and domains. The comparison against state-of-the-art methods aims to determine if symbolic systems can consistently extract accurate information from varied datasets or if there are limitations in their generalization capability that existing solutions can overcome.

Our system demonstrated stable performance across an extensive evaluation on four datasets from two distinct domains. It outperformed existing rule-based OIE systems on the `LSOIE` datasets, indicating that the learned patterns capture meaningful structural information. For the `WiRe57` dataset, it achieved a competitive balance between precision and recall, despite a slightly lower F_1 score compared to the best-performing state-of-the-art methods. The flexibility of our system allows users to adapt it to specific tasks by selecting between simple base patterns or more sophisticated ones, adjusting pattern and extraction counts, and applying a post-processing step after the initial tuple extraction.

The code of this work has been integrated into the author's fork[1] of `newpotato`[2], which is available open-source on GitHub and is released under the MIT License. The

[1] `https://github.com/whoopi24/newpotato` (Last accessed: April 17, 2025)
[2] `https://github.com/adaamko/newpotato` (Last accessed: April 17, 2025)

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

`newpotato` framework, named after `POTATO` [KGIR22], incorporates functionalities specifically designed for information extraction. Nevertheless, the actual implementation of an OIE system is part of this thesis. It is important to note that `newpotato` is still under development and should be considered an experimental system.

This thesis is structured as follows. Chapter 2 introduces the OIE task and its associated difficulties, and provides an overview of state-of-the-art methods and datasets in the field. Afterwards, basic information about the Semantic Hypergraph model and the `graphbrain` framework is given. Chapter 3 explains our novel supervised rule learning process, which builds upon the approach of Menezes and Roth [MR21] and incorporates annotations from `LSOIE` [SL21]. Chapter 4 describes the experimental setup of our OIE system, focusing on extraction strategies and providing examples to ensure the clarity and transparency of our method. The evaluation results are presented in Chapter 5, comparing our system's performance to state-of-the-art methods across four datasets from two different domains. In Chapter 6, limitations and potential improvements based on the evaluation and error analysis are discussed. Finally, Chapter 7 answers the research questions and summarizes the main contributions of this thesis.

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

CHAPTER 2

Background

This chapter begins with an introduction to the OIE task in Section 2.1, outlining its main challenges and reviewing state-of-the-art systems and benchmarks that have shaped the field, in order to provide context for the approach developed in this work. Afterwards, a detailed overview of the Semantic Hypergraph model, as proposed by Menezes and Roth [MR21], is given in Section 2.2.

2.1 The OIE Task

Open information extraction is the task of automatically extracting relational tuples from unstructured text without restricting extractions to a predefined set of relation types. It is considered a foundational preprocessing step for many downstream applications, such as question answering, text summarization and knowledge base construction.

Despite its importance and widespread use, there is no universal agreement on what constitutes a "correct" set of extractions. Consequently, multiple plausible tuples may exist for a given sentence, and determining a "complete" set of valid extractions is inherently difficult. This has led to inconsistencies across OIE systems, datasets, and evaluation methods, presenting persistent challenges for the OIE task:

1. **Varying Definitions:** Some OIE approaches extract only explicitly stated facts, while others include inferred or implicit relations, such as those conveyed through nominalizations or appositions. These conceptual differences influence both the design and the output of extraction systems.

2. **Dataset Discrepancies:** OIE datasets differ in their annotation guidelines and scope. While some focus strictly on surface-level predicates and arguments, others permit flexible argument structures (e.g. variable number of arguments) and include inferred content, making cross-dataset comparisons challenging.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3. **Evaluation Methods:** Standard NLP evaluations often rely on exact matching. However, the open-ended nature of OIE makes partial matches highly informative. Despite this, partial matching is not consistently supported by evaluation tools or adopted in benchmarks, complicating fair system comparisons.

Ultimately, some systems may perform better simply by producing more tuples, even if they are noisy and do not provide additional information. This is particularly problematic when measuring recall (how complete the extractions are). We illustrate one example of the `WiRe57` dataset. For the sentence "Chilly Gonzales (born Jason Charles Beck; 20 March 1972) is a Grammy-winning Canadian musician who resided in Paris, France for several years, and now lives in Cologne, Germany", 13 gold tuples are provided:

1. (Chilly Gonzales; [was] born; Jason Charles Beck)
2. (Chilly Gonzales; [was] born [on]; 20 March 1972)
3. (Chilly Gonzales; [was] born [in]; 1972)
4. (Chilly Gonzales; is; Canadian)
5. (Chilly Gonzales; is; a musician)
6. (Chilly Gonzales; is; a Grammy-winning Canadian musician)
7. (Chilly Gonzales; [won]; a [Grammy])
8. (who/(Chilly Gonzales); resided in; Paris, France; for several years)
9. (who/(Chilly Gonzales); resided in; France; for several years)
10. (Paris; [is in]; France)
11. (who/(Chilly Gonzales); lives in; Cologne, Germany; now)
12. (who/(Chilly Gonzales); lives in; Germany; now)
13. (Cologne; [is in]; Germany)

However, these extractions cannot be assumed to be complete. For instance, (Chilly Gonzales; [was] born; Jason Charles Beck; [on] 20 March 1972) could also be considered a valid extraction, combining the information given in tuples 1 and 2. Furthermore, a few tuples are highly similar, such as tuples 8 and 9 or 11 and 12, which include or omit city names in expressions of location. This highlights the nuanced and sometimes redundant nature of the annotations. Square brackets indicate inferred elements not explicitly stated in the original sentence. The predicates of tuples 7, 10 and 13 are solely inferred. Especially for rule-based systems, deriving such words is almost impossible, as they may struggle to recognize contextually implied elements.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

2.1.1 Rule-based Systems

Rule-based systems for OIE use hand-crafted or automatically learned patterns to identify and extract relational tuples from text. These rules typically operate on syntactic structures, allowing the system to capture meaningful relationships without requiring large amounts of labeled training data. In general, such systems are transparent, interpretable, and consistent, as they follow explicit logic, but they can be more rigid than data-driven neural methods when dealing with linguistic variation.

One of the first developments, TextRunner [YBB+07], introduces an unsupervised, rule-based system capable of handling large amounts of text data, laying the foundation for OIE research. ReVerb [FSE11] builds on TextRunner to improve precision by adding lexical and syntactic constraints on the relational part of the extractions. The open extractor `OLLIE` [MSS+12] refines these constraints and incorporates contextual information, expanding the syntactic scope to capture more complex relations.

Another approach is `ClausIE` [DCG13], which was developed by Del Corro and Gemulla and fundamentally differs from the aforementioned methods. Its unique feature is that it is clause-based, meaning it extracts tuples by identifying and processing individual clauses. Unlike statistical models, `ClausIE` relies on the linguistic knowledge of natural language, making it a state-of-the-art rule-based system suitable for our comparison. It uses a dependency parser to analyze syntactic structure of sentences (e.g., predicates and arguments), then splits sentences into individual clauses and assigns one of seven basic clause types to each of them. A key advantage is that the system does not require model training, making it independent of large datasets. In addition, `ClausIE` assigns confidence scores, obtained by the underlying dependency parser, to the extracted tuples. The system's performance heavily depends on the accuracy of this parser. `ClausIE` struggles with complex sentences containing multiple clauses connected by conjunctions like "and" or "or". It applies a decomposition process to simplify these coordinated conjunctions, but this often leads to incorrect parses. Since `ClausIE` focuses on syntactic correctness and has limited semantic understanding, essential adverbials may occasionally be omitted. `Stanford OIE` [AJPM15] extends the extraction of self-contained clauses by using a classifier-based approach, followed by triplet generation through a set of 14 hand-crafted patterns. Further advancements include `OpenIE4` [Mau16], `PropS` [SFDG16], and `MinIE` [GGdC17].

To systematically evaluate all of these emerging OIE systems, `WiRe57` [LGL19] was introduced as a benchmark dataset, enabling comparative analysis of OIE performance. The initial results can be found in Subsection 2.1.3 and the scoring procedure is further explained in Section 4.3. Menezes and Roth [MR21] developed a rule-based OIE system using semantic hypergraphs that outperformed all other evaluated systems on `WiRe57` at the time, reaching an F_1 score of 0.365 and recall of 32.6% with only five extraction patterns. This highlights the potential of rule-based models, which preserve the richness of natural language while providing a fully transparent and intelligible system.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

2.1.2 Neural Systems

Nevertheless, OIE research has increasingly shifted towards neural models to enhance both precision and scalability. As the first supervised learning method for OIE, RnnOIE [SMZD18] formulates OIE as a sequence labeling task using a Bidirectional Long Short-Term Memory (BiLSTM) trained on manually annotated data. It extends deep Semantic Role Labeling (SRL) models for triplet extraction and provides confidence scores, allowing for adjustments in the balance of precision and recall. RnnOIE marked an important shift from rule-based to neural methods, introducing greater flexibility but at the cost of reduced interpretability. Additionally, Stanovsky and Dagan [SD16] introduced the first large-scale, independently constructed OIE corpus by automatically converting question-answer driven Semantic Role Labeling (QA-SRL) annotations [HLZ15] into OIE tuples. In the literature, it is often referred to as OIE2016. More information on OIE datasets follows in Subsection 2.1.3.

SpanOIE [ZZ20] represents the first attempt to frame OIE as a span-based prediction task. The model predicts the boundaries of arguments and relations using a neural network, outperforming previous approaches on the OIE2016 dataset and its re-annotated version, Re-OIE2016. However, a key limitation is that it cannot capture the broader context of a sentence. In parallel, Kolluru et al. developed IMoJIE [KAR+20], a sequence generation model that produces the next extraction conditioned on all previously extracted tuples. This allows the model to generate a variable number of diverse extractions per sentence, addressing the limitations of its underlying model, CopyAttention [CWZ18]. IMoJIE showed a significant improvement in extraction quality and influenced the development of subsequent models. For instance, OpenIE6 [KAA+20] introduces a novel iterative labeling-based architecture, Iterative Grid Labeling (IGL), which treats OIE as a 2-D grid labeling task. It also incorporates a newly designed coordination analyzer capable of handling conjunctive sentences. OpenIE6 outperforms previous models on the CaRB [BAM19] benchmark and features a highly efficient pipeline that is ten times faster than the state-of-the-art, making it one of the most competitive OIE systems available today.

Alternatively, MacroIE is a non-autoregressive OIE system designed to overcome a key limitation of existing approaches, which typically extract facts sequentially by predicting each new fact based on the previously decoded ones. This enforces an unnecessary order and leads to error accumulation across subsequent steps. MacroIE avoids this bottleneck by constructing a graph of fact elements and identifying maximal cliques, enabling the simultaneous extraction of multiple facts. The modular and iterative multilingual model MILIE [KGR+22] extracts the different slots of a tuple iteratively using a BERT-based transformer as its core architecture. As a hybrid system, it combines neural and rule-based approaches, allowing the rule-based component to compensate for the lack of exhaustive training data. MILIE is evaluated on multilingual text corpora, including English, Chinese, German, Arabic, Galician, Spanish, and Portuguese, where it significantly outperforms other neural baselines on the BenchIE [GYK+22] benchmark. However, its performance remains below rule-based systems like ClausIE and MinIE. A key strength of MILIE lies in its ability to handle multiple languages besides English effectively.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

2.1.3 Datasets

Alongside the deployment of OIE systems, the creation of suitable evaluation datasets has been essential. This subsection provides a more detailed overview of the datasets mentioned above, particularly focusing on those used to train and test the system developed in this thesis.

The first independent and large-scale OIE corpus, `OIE2016` [SD16], was generated by automatically converting `QA-SRL` annotations [HLZ15] into OIE tuples. It was specifically designed to serve as a benchmark for evaluating OIE systems, facilitating the development of supervised OIE approaches. In total, it comprises 10,359 extractions from 3,200 sentences, originally sourced from `Wikipedia` and newswire articles. A comparison of `OpenIE4`, `ClausIE`, `OLLIE`, `PropS`, `Stanford OIE`, and ReVerb reveals the following: `OpenIE4` achieves the highest precision (above 78%) among all systems for recall levels above 3%, and also obtains the best area under the curve (AUC) score. `ClausIE` leads in recall performance, reaching over 81%. `Stanford OIE` assigns a confidence score of 1 to 94% of its extractions, which indicates overconfidence and helps explain its low precision, as it fails to distinguish between correct and incorrect outputs.

`CaRB` (short for "Crowdsourced Automatic Open Relation Extraction Benchmark"), introduced by Bhardwaj et al. in 2019 [BAM19], was created by re-annotating the dev and test splits of the `OIE2016` dataset through crowdsourcing. The goal was to improve annotation quality and provide a more reliable evaluation framework for OIE compared to `OIE2016`, which was confirmed by NLP experts. On this benchmark, neural models currently lead the field. `MacroIE` achieves the highest F_1 score of 0.548, followed closely by `IMoJIE` with 0.535, though it is relatively inefficient, processing only 2.6 sentences per second. In third place is `OpenIE6`, scoring 0.527 while handling 31.7 sentences per second. The best-performing rule-based system is `OpenIE4`, with an F_1 score of 0.516 and a throughput of 20.1 sentences per second.

The `WiRe57` dataset, published by Léchelle, Gotti, and Langlais [LGL19], was initially used to evaluate the performance of seven rule-based OIE systems. The dataset consists of 57 sentences extracted from `Wikipedia` and the news agency `Reuters`, comprising 343 high-quality annotations. Of these, 57% contain anaphoras and 54% include inferred words that are not part of the original sentence. Most tuples (74%) represent binary relations. The accompanying scorer supports both exact and partial matching. Detailed information on the scoring procedure is provided in Section 4.3.

A comparison of all rule-based systems mentioned in Subsection 2.1.1, excluding TextRunner, showed that `MinIE` performed best regarding F_1 score (0.358) and recall (32.3%), closely followed by `ClausIE` with an F_1 score of 0.342. ReVerb led in precision (56.9%) but at the cost of lowest recall (12.1%). Menezes and Roth [MR21] achieved an F_1 score of 0.365 and a recall of 32.6% using only five extraction patterns with their semantic hypergraph-based approach, outperforming all previously reported results. This makes `WiRe57` a suitable benchmark for comparing our system with state-of-the-art methods and for addressing RQ2.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Another important OIE dataset for our work, LSOIE, was presented by Solawetz and Larson [SL21] as a large-scale resource for supervised OIE. It is derived from the QA-SRL 2.0 dataset [FMHZ18], following a similar conversion process as used for OIE2016. It is known for its large size compared to other human-annotated OIE datasets and comprises two corpora, LSOIE-wiki and LSOIE-sci, which distinguish between the domains of Wikipedia articles and scientific texts. These are further split into train, dev and test sets. The creators also constructed and evaluated several benchmark OIE models on LSOIE, establishing baselines for future progress on the task. All models are based on a replication of RnnOIE [SMZD18], framing the OIE task as BIO tagging with tunable extraction thresholding. On LSOIE-wiki/test, an F_1 score of 0.31 was achieved, while LSOIE-sci/test reached an F_1 score of 0.38.

For developing the rules for our system, we focus on the training part of the LSOIE-wiki dataset, since Menezes and Roth [MR21] also worked with text from Wikipedia for their unsupervised OIE approach. It contains 46,015 tuples from 19,675 sentences, which are annotated at the token level with P (for predicate) and A_0 to A_N (for arguments). Tokens marked as O are not extracted. The dataset focuses on explicit extractions, where relations are directly stated by verbal predicates rather than implicitly through nominalizations. The annotations include gold tuples without objects (consisting only of argument A_0 and predicate P), which does not align with our objectives. Therefore, we take additional actions to exclude these cases. We test our system with the two remaining parts of the LSOIE-wiki dataset, as well as the test set of LSOIE-sci.

Unlike existing OIE benchmarks, BenchIE adopts a fact-based evaluation approach that accounts for the informational equivalence of extractions across three languages: English, Chinese and German. Its gold standard, based on 300 sentences of CaRB [BAM19], is organized into fact synsets, which are clusters that comprehensively enumerate all acceptable surface forms of the same fact. To better reflect common downstream applications, BenchIE is also multi-faceted, offering variants that emphasize different aspects of OIE evaluation, such as the compactness or minimality of extractions. Evaluations of several state-of-the-art OIE systems on BenchIE show that their effectiveness is significantly lower than indicated by CaRB. Notably, neural models underperform compared to rule-based approaches. ClausIE and MinIE lead the leaderboard with an F_1 score of 0.34, followed by OpenIE6 at 0.25. ClausIE also achieves the highest precision (50%), while MinIE attains the highest recall (28%).

An overview of all datasets used in this thesis is illustrated in Table 2.1. We chose not to use further datasets for evaluation to avoid the influence of too many different annotators. We believe that the domain switch within the LSOIE datasets is sufficient to answer RQ3, which addresses generalizability across various datasets. We build on existing OIE approaches, specifically leveraging semantic hypergraphs to enhance interpretability and consistency. We focus on a supervised learning approach, incorporating the annotations provided by LSOIE to develop a transparent, rule-based system that facilitates user understanding and trust.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Dataset	Domains	#Sentences	#Extractions
LSOIE-sci/test	Science	9,219	17,001
LSOIE-wiki/train	Wiki, Wikinews	19,675	46,015
LSOIE-wiki/dev	Wiki, Wikinews	2,266	5,268
LSOIE-wiki/test	Wiki, Wikinews	2,403	5,373
WiRe57	Wikipedia, Newswire	57	343

Table 2.1: Overview of datasets with sentence and extraction counts.

2.2 Semantic Hypergraphs

The concept of the Semantic Hypergraph (SH) was developed by Menezes and Roth [MR21] and serves as a powerful knowledge model. Its main functionality is to represent natural language while preserving its hierarchical structure and complexity. It can be described as a hybrid approach, leveraging the benefits of machine learning and symbolic methods while also minimizing ambiguity and structural variability. This section gives a brief overview of semantic hypergraphs, starting with their formal definition, followed by their notation used in the literature, and concluding with their usage in a variety of tasks. Menezes and Roth [MR21] created an open-source software library called `graphbrain`, which provides a comprehensive set of tools for constructing, manipulating, and analyzing semantic hypergraphs. Subsection 2.2.4 presents the key features of `graphbrain` that are utilized in this thesis.

2.2.1 Definition

Traditional graph-based and distributional methods, which represent words based on their contextual usage, struggle to fully capture the complexity of natural language. Natural language is recursive, allowing concepts to be built from other concepts, and it is capable of expressing multi-way (n-ary) relationships. These two properties motivate the following definition of the Semantic Hypergraph:

A Semantic Hypergraph $H = (V, E)$ consists of a vertex set V and an edge set E which contains so-called hyperedges $(e_i)_{i \in 1 \dots M}$, each connecting an arbitrary number of vertices. Furthermore, hyperedges have two key characteristics: order and recursivity. Order emphasizes the importance of the specific vertex positions within a hyperedge, which is similar to the concept of directed graphs. Recursivity allows hyperedges to occur as vertices within other hyperedges, enabling the representation of higher-order relationships. Atomic hyperedges, meaning irreducible hyperedges of size one, are denoted as “atoms”. In the following, the term “hyperedges” will encompass all types of hyperedges, including atomic and non-atomic structures. Thus, the corresponding hypergraph will be referred to as a “semantic hypergraph”.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

2.2.2 Notation

To comprehend the ideas and the novel rule learning process described in this thesis, it is necessary to introduce a few SH-related terms. However, for a more in-depth perspective, readers are encouraged to consult the original paper and the manual available on the `graphbrain` website[1].

As emphasized by Menezes and Roth [MR21], the SH notation involves two principles:

1. Every hyperedge belongs to one of eight fundamental types.
2. Every hyperedge starts with a connector, followed by arguments, which can themselves be either atomic or non-atomic hyperedges.

The structure of a hyperedge effectively captures the meaning of a natural language sentence. Among the eight hyperedge types, depicted in Table 2.2, the first six can be explicit, meaning they directly annotate an atomic hyperedge, or implicit, meaning they are derived from the types of the hyperedge's elements. The last two types are always implicit, as they emerge from the composition of hyperedges. At this point, we want to mention two special atoms. The special builder atom, denoted by `+/B`, defines compound nouns, while `:/J` represents the generic conjunction.

Code	Type	Purpose	Example
Atomic or non-atomic			
C	Concept	Define atomic concepts	apple/C
P	Predicate	Build relations	(is/P berlin/C nice/C)
M	Modifier	Modify any other hyperedge type, including itself	(red/M shoes/C)
B	Builder	Build concepts from concepts	(of/B capital/C germany/C)
T	Trigger	Build specifications	(in/T 1994/C)
J	Conjunction	Define sequences of hyperedges	(and/J meat/C potatoes/C)
Non-atomic only			
R	Relation	Express facts, statements, questions, orders, ...	(is/P berlin/C nice/C)
S	Specifier	Relation specification (e.g., condition, time, ...)	(in/T 1976/C)

Table 2.2: Overview of semantic hyperedge types [MR21].

The type of a non-atomic hyperedge is implicit and can be inferred from its connector type and arguments, following the rules shown in Table 2.3. Regular expression notation

[1] `https://graphbrain.net` (Last accessed: April 17, 2025)

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

is used here: the symbol + denotes one or more occurrences of the preceding entity type, square brackets indicate multiple possible options, and `x` and `y` represent any type.

Element types →	**Resulting type**
(M x)	x
(B C C+)	C
(T [CR])	S
(P [CRS]+)	P
(J x y+)	x

Table 2.3: Type inference rules [MR21].

"Argument roles" are another construct that describes the types of specific hyperedge components, namely builders and predicates. They are encoded as character sequences that specify the role of each argument in relation to the corresponding connector. For builders, we differentiate between main concepts (m) and auxiliary concepts (a). In contrast, predicates have a much broader range of roles, as shown in Table 2.4, with the top six being the most common. Argument roles in relations are necessary when the role cannot be obtained solely from its type. For instance, the same concept may appear either as a subject or as an object in a relation.

Role	**Code**
active subject	s
passive subject	p
agent (passive)	a
subject complement	c
direct object	o
indirect object	i
parataxis	t
interjection	j
specification	x
relative relation	r

Table 2.4: Predicate argument roles [MR21].

We explain the concept of "functional pattern expressions", since they play an essential role in our supervised learning approach. Their general form resembles function applications in `LISP`-like languages. We make use of the `var` functional pattern, which enables specifying a segment of a pattern and storing it as a variable. This allows capturing complex parts of a semantic hyperedge for further usage. It has the general form: `(var pattern-edge variable-name)`. From this point forward, the term "subedges" refers to all components of a hyperedge, including both atomic and non-atomic structures. Moreover, the "root" of an atom denotes its fundamental base form, representing its core meaning while ignoring types and argument roles. It is the essential element that links an atom to its primary concept, for instance, `has` is the root of the atom `has/P.so`.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

2.2.3 Applications

The SH model has a wide range of applications in natural language processing and knowledge representation. It can be used for identifying linguistic structures with the help of symbolic patterns. A key use case is open information extraction, where hypergraphs facilitate knowledge extraction from unstructured text by capturing relationships between entities in a flexible and expressive manner. Beyond OIE, semantic hypergraphs are useful for conjunction decomposition, concept taxonomy inference, and coreference resolution. Coreference resolution is the task of identifying all words or phrases in a text that refer to the same real-world entity. For example, in the sentence "Mia lost her keys, but she found them later", "she" refers to "Mia", and "them" refers to "her keys". A coreference resolution system identifies these links so that the text makes sense as a whole. Accurate coreference resolution is essential for many downstream tasks in NLP. Menezes and Roth [MR21] also presented a case study of a claim and conflict analysis, where statements made by different actors can be linked, analyzed, and inferred symbolically within the hypergraph structure.

2.2.4 `graphbrain` Library

In this subsection, we outline the key functionalities of `graphbrain`[2] that were used in this thesis. The open-source Python library `graphbrain` is designed for constructing, modifying, and analyzing hypergraphs. It facilitates automated meaning extraction and text comprehension, while also supporting knowledge exploration and inference.

A fundamental component is `create_parser()`[3], which converts sentences into hyperedge structures. Since our supervised rule learning approach focuses on semantic hyperedges, we rely on several functions from the `Hyperedge` class[4]. These include:

- `hedge()`, which constructs a `Hyperedge` object from a list, tuple, or string,
- `mtype()`, which determines the main type of an edge after type inference,
- `argroles()`, which retrieves the argument roles of an edge, and
- `atoms()`, which lists all unique atoms contained in the edge.

Additionally, we make use of functions related to the `var` functional pattern, such as `apply_variables()`, `contains_variable()`, and `all_variables()`[5]. For the

[2] `https://github.com/graphbrain/graphbrain` (Last accessed: April 17, 2025)

[3] `https://github.com/graphbrain/graphbrain/blob/master/graphbrain/parsers` (Last accessed: April 17, 2025)

[4] `https://github.com/graphbrain/graphbrain/blob/master/graphbrain/hyperedge.pyx` (Last accessed: April 17, 2025)

[5] `https://github.com/graphbrain/graphbrain/blob/master/graphbrain/patterns/variables.pyx` (Last accessed: April 17, 2025)

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

tuple extraction, we leverage the conjunction decomposition functionality[6], which breaks down complex hyperedges into their individual parts. Finally, we apply the `Matcher` class, specifically the `match_pattern(edge, pattern)` function, to match the decomposed edges with our learned rules. As previously mentioned, this section does not provide a complete list of all `graphbrain` features used in our work, but highlights the most relevant ones.

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

[6] `https://github.com/graphbrain/graphbrain/blob/master/graphbrain/utils/conjunctions.py` (Last accessed: April 17, 2025)

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

CHAPTER 3

Method

In this chapter, we present our supervised rule learning method using semantic hyperedges in combination with `LSOIE` annotations. The initial idea was to follow a similar OIE approach as described by Menezes and Roth [MR21]. Their system achieves promising results, as it outperforms the state-of-the-art on the `WiRe57` dataset in terms of F_1 score and recall. However, they use an unsupervised learning approach to generate rules, which requires manual effort. To address this limitation, this thesis focuses on experiments with different strategies, primarily supervised learning, to develop a more efficient but still effective OIE system. Our key innovation is the integration of annotations from the `LSOIE-wiki/train` dataset into the rule learning process. This chapter outlines the main concepts behind our method and details its technical implementation, including all intermediate steps required to convert labeled sentences into abstract symbolic patterns suitable for OIE.

As noted in Subsection 2.2.4, Menezes and Roth developed `graphbrain`[1], which provides an extensive set of tools for handling semantic hypergraphs. However, their OIE system itself is not publicly available. It was challenging to directly build upon their approach, as we had to reconstruct their system solely based on the details provided in their paper. At a certain stage, we realized that shifting from an unsupervised to a supervised learning strategy was favorable. This required a precise analysis of the existing `graphbrain` functionalities in order to determine which components could be adapted for our purposes and which needed to be implemented from scratch.

In the following, the rule learning process is illustrated using the example sentence: "The controversial investments were made between 2007 and 2009, Panorama explains". The first step is to parse the sentence with the `create_parser()` function from the `graphbrain` library. Figure 3.1 shows the resulting semantic hyperedge. As defined by Menezes and Roth [MR21], a semantic hyperedge "reflects the meaning of the sentence

[1] `https://graphbrain.net` (Last accessed: April 17, 2025)

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

and respects the SH syntactic rules". Hereafter, we always refer to semantic hyperedges when hyperedges are mentioned. As noted in Section 2.2, hyperedges follow a specific structure that reflects the syntax of natural language. As a result, complex sentences produce nested hyperedges, which can complicate the succeeding generalization process. This nesting is visualized through a block structure, as shown in Figure 3.1. For example, while "explains" and "panorama" are represented as individual atoms, the remainder of the sentence forms a subedge that includes three components: `(were/M made/P.px)`, `(the/M (controversial/M investments/C))`, and `(between/T (and/J 2007/C 2009/C))`, the latter two of which are further nested.

Sentence:
The controversial investments were made between 2007 and 2009, Panorama explains.

Semantic Hyperedge:

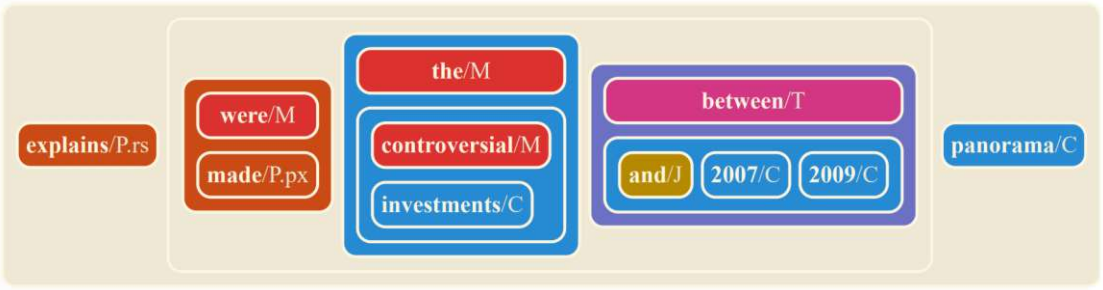


Figure 3.1: Example sentence with its corresponding semantic hyperedge.

Index	Token	Triplet 1	Triplet 2
0	The	A1-B	A0-B
1	controversial	A1-I	A0-I
2	investments	A1-I	A0-I
3	were	A1-I	O
4	made	A1-I	P-B
5	between	A1-I	A1-B
6	2007	A1-I	A1-I
7	and	A1-I	A1-I
8	2009	A1-I	A1-I
9	,	O	O
10	Panorama	A0-B	O
11	explains	P-B	O
12	.	O	O

Table 3.1: Examples of token-based `LSOIE` annotations.

The next step is to leverage the `LSOIE` annotations to automate and accelerate the rule

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

learning process. Each token of a sentence is labeled with tags: P (for predicate), A_0 to A_N (for arguments), and O (non-extraction tokens). An additional marker $-B$ indicates the beginning of an annotation span, while $-I$ marks its continuation. Since semantic hyperedges are also token-based, this alignment makes it possible to identify labeled spans within the hypergraph structure. Table 3.1 shows two gold-standard tuples for our example sentence, preserving the structure of the original `LSOIE` annotations.

We map the annotations to their corresponding hyperedges with the goal of identifying exact matches. For each annotation label (e.g., A_0, P, etc.), we search for a subedge that *exclusively* contains all tokens assigned to that label. In other words, each annotation span must be fully covered by a single atom or subedge, without including any additional tokens. This condition ensures clean and precise mappings. Table 3.2 provides the successful mappings of both triplets for our example sentence.

Variable	**Mapped Triplet 1**	**Mapped Triplet 2**
ARG0	panorama/C	(the/M (controversial/M investments/C))
REL	explains/P.rs	made/P.px
ARG1	((were/M made/P.px) (the/M (controversial/M investments/C)) (between/T (and/J 2007/C 2009/C)))	(between/T (and/J 2007/C 2009/C))

Table 3.2: Successful mappings of annotations to parts of the semantic hyperedge.

Figure 3.2 illustrates that due to the nested structure of semantic hyperedges, exact matches are not always possible. In this example, the last argument "on February 2" should be mapped to the subedge `(on/T (of/B.ma (+/B.ma february/C 2/C) (next/M year/C)))`, which also includes the tokens "of next year." Since this subedge cannot be further decomposed in a way that isolates only the relevant tokens, no exact match can be established. As a result, the corresponding tuple is excluded from the rule learning process.

Referring to the example presented in Figure 3.2, one might argue that the `LSOIE` annotation is flawed, as "on February 2" is an incomplete representation of the full expression "on February 2 of next year". This highlights that our system is able to detect potential annotation inconsistencies, which should be revised to improve dataset quality. To enforce that annotations must align with subedges from the semantic hyperedge, we used the `map_to_subgraphs()` function of the `GraphbrainExtractor` from `newpotato` with the `strict` argument set to `True`. The parsing and mapping procedure constitutes the time-consuming part of the rule learning process, taking around 80 minutes for the entire `LSOIE-wiki/train` dataset.

Further on, the annotations are incorporated into the semantic hyperedge structure as functional patterns by using the `apply_variables()` function from `graphbrain`.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Sentence: The new elections are scheduled to take place on February 2 of next year.

LSOIE Annotation: (The new elections; scheduled; to take place; on February 2)

Semantic Hyperedge:

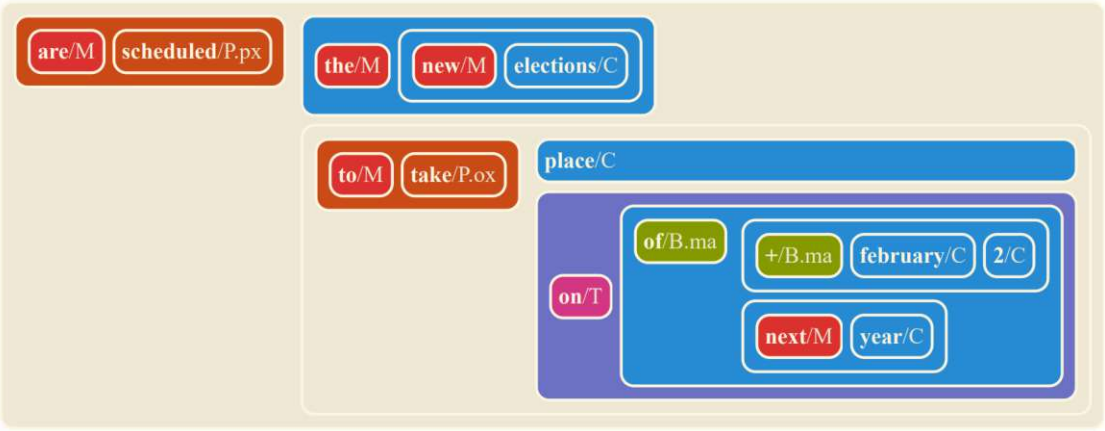


Figure 3.2: Failed mapping of argument “on February 2” to semantic hyperedge.

Sentence:
The controversial investments were made between 2007 and 2009, Panorama explains.

LSOIE Annotation:
(Panorama; explains; the controversial investments were made between 2007 and 2009)

Semantic Hyperedge with Functional Patterns:

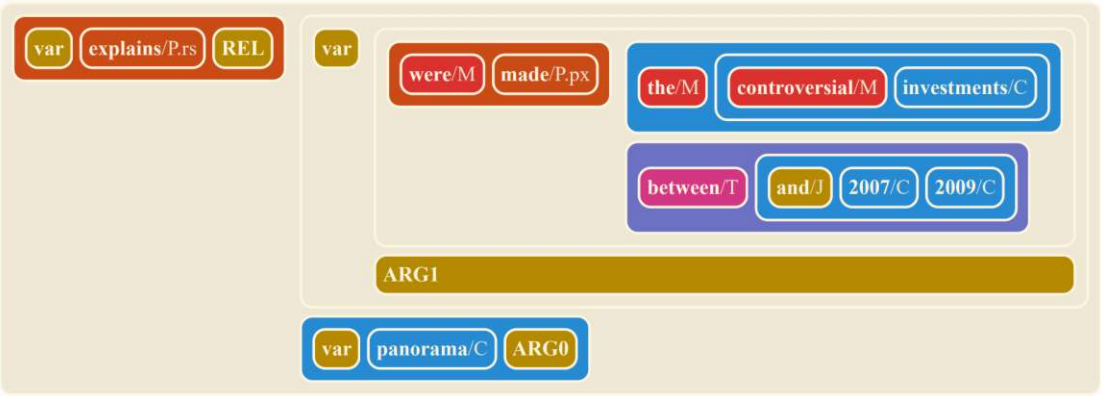


Figure 3.3: Successful integration of annotations with functional patterns.

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

These constructs enclose specific parts of a semantic hyperedge to mark individual annotation labels. More precisely, each mapped atom or subedge is surrounded by parentheses, the term `var` and the annotation label, as illustrated in Figure 3.3.

To better explain our generalization method, we first summarize the simplification procedure used by Menezes and Roth [MR21]: Each subedge of the hyperedge is replaced with *, referred to as a "wildcard", while retaining its inferred type information. For example, the following transformation is demonstrated in their paper:

`(is/P.sc aragorn/C (of/B.ma king/C gondor/C))` → `(*/P.sc */C */C)`

After performing type inference on the subedge `(of/B.ma king/C gondor/C)`, it transforms into a wildcard from type concept `(*/C)`. This abstraction may be recursively applied, expanding subedges while adhering to the type inference rules outlined in Table 2.3 in reverse order, such as:

`(*/P.sc */C */C)` → `(*/P.sc */C (*/B.ma */C */C))`

In our approach, the key abstraction step involves removing the content enclosed by each functional pattern and replacing it with the corresponding annotation label. We illustrate this step using our ongoing example, as shown in Figure 3.4. Every component of this hyperedge represents a functional pattern, such as `(var explains/P.rs REL)`. During abstraction, the token "explains" is removed and replaced with the annotation label `REL`, while the type and argument roles `/P.rs` are retained, resulting in `REL/P.rs`. This transformation is repeated for each annotated part of the hyperedge. Each unannotated component, including atomic and non-atomic elements, is replaced by a basic wildcard, which matches any hyperedge. To preserve type information, the main type of the eliminated parts may be inferred, conformable to Table 2.3. For predicates and builders, the argument roles, specifying the relation to their corresponding arguments, follow.

We allow to expand subedges to the next level to prevent overgeneralization. This feature is optional and generates additional, more sophisticated patterns. Figure 3.5 demonstrates this procedure for our example sentence. Due to the nested structure, the second component of the semantic hyperedge contains three functional patterns with different variables. Expanding the subedge to its next level means that we decompose it one level deeper and treat the inner subedges as individual components. At this point, the annotation labels are isolated, enabling the abstraction step. This expansion is also applied to unannotated parts of the hyperedge, following the transformation process proposed by Menezes and Roth [MR21].

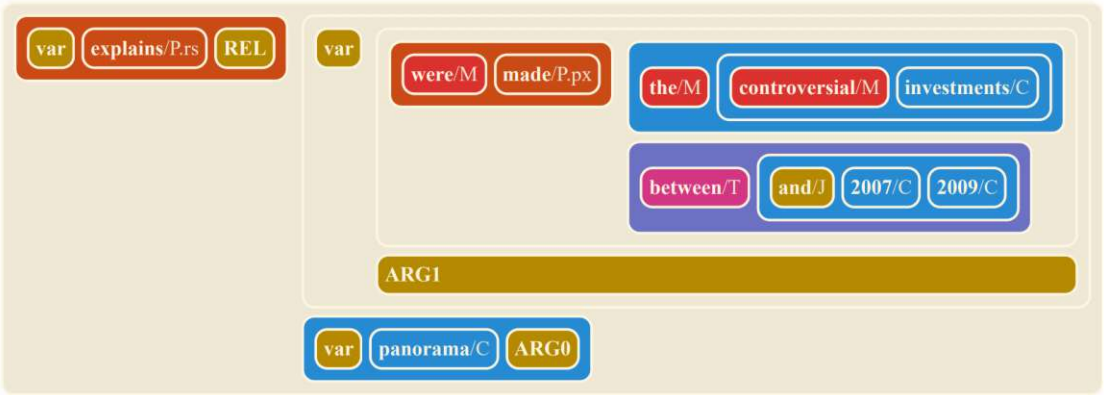
As depicted in Figure 3.5, the generalization process produces the pattern `(*/P.rs (REL/P.px ARG0/C ARG1/S) */C)`. We ensure that each pattern contains at least three annotation labels, thereby filtering out patterns that lack an object, which is possible due to certain `LSOIE` annotations. In our example, the subpattern `(REL/P.px`

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Sentence:
The controversial investments were made between 2007 and 2009, Panorama explains.

LSOIE Annotation:
(Panorama; explains; the controversial investments were made between 2007 and 2009)

Semantic Hyperedge with Functional Patterns:



Generalization Process:

`(var explains/P.rs REL)`	→ **REL/P.rs**
`(var ((were/M made/P.px) (the/M [...] )) ARG1)`	→ **ARG1/R**
`(var panorama/C ARG0)`	→ **ARG0/C**

Final Pattern: `(REL/P.rs ARG1/R ARG0/C)`

Figure 3.4: Successful integration of annotations and generalization process (Example 1).

`ARG0/C ARG1/S)` already includes all annotation labels and is therefore valid as well. All resulting patterns are collected in a `Counter` object from the `collections` Python package. In the end, patterns are sorted in descending order of frequency, and the top N patterns are returned, depending on the user's choice. Table 3.3 summarizes the main steps of our supervised learning approach along with the corresponding functions from the `newpotato` framework.

#	Step	`newpotato` Functions
1	Text Parsing	`get_graphs()`
2	Triplet Mapping	`map_triplet()`, `map_to_subgraphs()`
3	Annotation Integration	`get_annotated_sentences()`
4	Generalization Process	`generalize_edge()`, `edge2pattern()`
5	Pattern Counting	`extract_rules()`, `get_rules()`

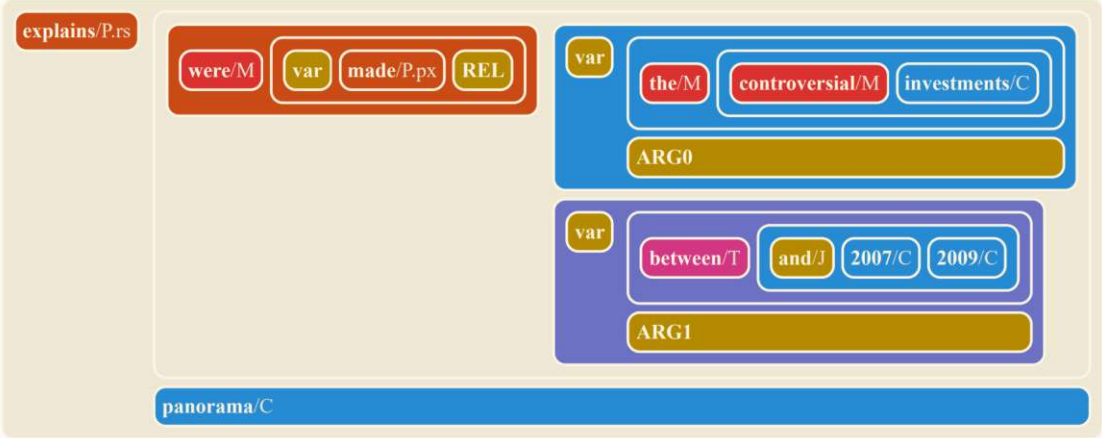
Table 3.3: Overview of rule learning steps and corresponding functions.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Sentence:
The controversial investments were made between 2007 and 2009, Panorama explains.

LSOIE Annotation: (The controversial investments; made; between 2007 and 2009)

Semantic Hyperedge with Functional Patterns:



Generalization Process (with expanded subedges):

`explains/Pd.rs`	→ ***/P.rs**
`(were/M (var made/P.px REL))`	→ **REL/P.px**
`(var (the/M (controversial/M investments/C)) ARG0)`	→ **ARG0/C**
`(var (between/T (and/J 2007/C 2009/C)) ARG1)`	→ **ARG1/S**
`panorama/C`	→ ***/C**

Final Patterns:

```
(*/P.rs (REL/P.px ARG0/C ARG1/S) */C)
(REL/P.px ARG0/C ARG1/S)
```

Figure 3.5: Successful integration of annotations and generalization process (Example 2).

Menezes and Roth [MR21] limited their analysis to the 50 most frequent patterns, manually converting them into valid OIE relations. These were further compressed using regular expressions, resulting in the five final patterns shown in Table 3.4. These five patterns can be decompressed into a full set of 13 rules, including optional arguments (denoted with three dots).

Table 3.5 presents a compressed version of our top 10 patterns, whose structure closely mirrors the final patterns shown in Table 3.4. Notably, all of our patterns resemble the first pattern from the original approach. This pattern begins with the relation `REL` of type predicate (`/P`), followed by argument `ARG1` of type concept (`/C`), then `ARG2`, and optionally `ARG3`. However, switching to a supervised learning approach led to minor differences in the pattern format. In particular, split relations (e.g., `REL1`, `REL2`, or

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

REL3) are absent, since LSOIE represents relations as unified spans within a single annotation (denoted with REL). Additionally, our method does not produce patterns involving the special builder atom +/B. In the original work, the relation for such patterns is assumed to be "is" during tuple extraction. This omission means that simple inferences involving compound nouns may be overlooked.

#	Pattern
1	(REL/P.{[sp][cora]x} ARG1/C ARG2 ARG3...)
2	(+/B.{m[ma]} (ARG1/C...) (ARG2/C...))
3	(REL1/P.{sx}-oc ARG1/C (REL2/T ARG2))
4	(REL1/P.{px} ARG1/C (REL2/T ARG2))
5	(REL1/P.{sc} ARG1/C (REL3/B REL2/C ARG2/C))

Table 3.4: Final patterns from original SH approach [MR21].

#	Pattern	#Cases
1	(REL/P.{[sp][rx]} ARG0/C ARG1/[RS])	1,107
2	(REL/P.{so} ARG0/C ARG1/C)	475
3	(REL/P.{sox} ARG0/C ARG1/C */[RS])	150
4	(REL/P.{sox} ARG0/C ARG1/C ARG2/S)	124

Table 3.5: Top 10 compressed patterns with number of cases.

To conclude, we integrated token-based annotations from the LSOIE dataset into semantic hyperedges to create a novel supervised learning approach that automates the rule learning process and eliminates the need for manual pattern identification. While our approach results in patterns similar to those from the original OIE system, they differ in key aspects, such as the absence of split relations and the special builder atom +/B, which affects the handling of compound nouns. Notably, all of the top ten patterns resemble the first pattern from the original method. This lays the foundation for the next chapter, which presents the experimental setup of our work, including the tuple extraction process.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

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

CHAPTER 4

Experimental Setup

This chapter outlines the experimental setup for evaluating our OIE system. First, we describe how the tuple extraction is conducted. To answer RQ1, we experimented with various key configurations to identify the optimal setup for our OIE system. The so-called `WiRe57` scorer is used as evaluation method, which offers a standardized framework for assessing OIE system performance, as detailed in Section 4.3. This section also introduces the evaluation metrics used to measure extraction quality, including precision, recall, and F_1 score. Finally, the comparison systems are described.

4.1 Tuple Extraction and Gold Data Reduction

4.1.1 Initial Tuple Extraction

After developing the rules with the training text corpus, they will be evaluated on unseen data. Ultimately, the patterns will be used for extracting tuples, which will be compared to the ground truth of a test dataset. Each sentence undergoes the following procedure:

Initially, the input sentence is parsed into its semantic hyperedge structure. If an error occurs during the parsing process, the sentence is not taken into consideration for the evaluation. Then, conjunction decomposition is performed according to the `graphbrain` function `conjunctions_decomposition(edge, concepts=True)`. Every part of the decomposed hyperedge is matched against the list of rules chosen for the evaluation. We start with the first, most common pattern and continue to find matches with the subsequent elements of the pattern set until the maximum number of extractions or the end of the list is reached. For matching the pattern to the specific hyperedge, the function `match_pattern()` of the `graphbrain` library is applied.

4.1.2 Gold Data Reduction

After processing each sentence and extracting the initial tuples, the gold tuples are stored in a `JSON` file formatted for compatibility with the `WiRe57` scorer. Approximately 20% of the extractions in each `LSOIE` dataset are annotated without an object. Since our evaluation framework does not support such cases (see Section 4.3) and they are not the focus of our work, we excluded all sentences containing at least one such annotation. Table 4.1 shows that this results in a substantial reduction in dataset size, however, it ensures a consistent evaluation.

	Before		After	
Dataset	**#Sent**	**#Extr**	**#Sent**	**#Extr**
`LSOIE-sci/test`	9,219	17,001	6,116	9,901
`LSOIE-wiki/dev`	2,266	5,268	1,483	2,932
`LSOIE-wiki/test`	2,403	5,373	1,567	2,983

Table 4.1: Sizes of `LSOIE` datasets before and after filtering.

4.1.3 Post-processing of Extractions

We noticed that a high amount of tuples with merely varying arguments is extracted. To reduce noise and increase precision, we added a post-processing step that combines extractions from the same sentence with identical predicates and first arguments. The second argument is taken from the first extraction in its original position. Subsequently, any additional arguments from the current tuple, as well as those from similar tuples, are included as long as they provide new information.

We implemented the function `is_similar(s1, s2, threshold=0.5)`, which compares pairs of arguments, `s1` and `s2`, by using string similarity to identify redundant objects. It uses the `SequenceMatcher` from the `difflib` package, which calculates a similarity score (accessible with `ratio()`) between strings. If the similarity score is greater than or equal to the specified threshold, the strings are considered similar. We only keep arguments that are sufficiently different (i.e., their similarity score is below 0.5). In the end, the combined extractions are saved in the same format as the gold tuples. This step can be omitted by setting the argument `comb` in our implemented function `generate_triplets_and_gold_data()` to `False`. We refer to this scenario of omission as **nc** (for "no combination") in Chapter 5 and demonstrate the process through an example:

Consider the sentence 'He went on to say in the same post, "Nelson Mandela did great things for his country and was a brave man but he was not an AMERICAN!!!"'. Table 4.2 shows four similar extractions with subtle differences in the second and third arguments. Notably, tuples #3 and #4 are combinations of #1 and #2, generated by different symbolic patterns. #4 also presents the arguments in a different order compared to #3.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Since all four extraction candidates share the same first argument and relation, the post-processing step can be applied. We begin with extraction #1, which is taken as is. Next, the argument "for his country" from tuple #2 is compared to the existing arguments (excluding the identical first argument), in this case "great things", using the string similarity function described earlier. As the similarity score is below 0.5, "for his country" is added as a third argument to the final extraction.

Tuple #3 is then compared pairwise with the current extraction arguments. Since it does not contain additional information, it is disregarded. The same applies to tuple #4. In the end, we obtain the single consolidated extraction (Nelson Mandela; did; great things; for his country), instead of four similar ones.

Variable	**Extr. #1**	**Extr. #2**	**Extr. #3**	**Extr. #4**
arg1	Nelson Mandela	Nelson Mandela	Nelson Mandela	Nelson Mandela
rel	did	did	did	did
arg2	great things	for his country	great things	for his country
arg3	–	–	for his country	great things

Table 4.2: Four similar extractions are combined into a single tuple: (Nelson Mandela; did; great things; for his country).

4.2 Key Configurations

The intention of this thesis is to build on the OIE approach of Menezes and Roth [MR21]. RQ1 is about the different strategies that can be used to find effective patterns for the OIE task. In the early stages of our work, we decided to use a supervised learning approach to avoid a manual inspection of the rules and to make the process more efficient. Nevertheless, we had to make additional design choices during the development of our OIE system, which we want to summarize in the following section.

Regarding the rule learning process, we adopted the restriction mentioned by Menezes and Roth [MR21] that only relations of size 3 or 4 are used. They argue that smaller relations cannot contain triplets, while larger ones might embed the triplet in a smaller subgraph that meets this condition. We tested our system without these restrictions (and without expanded subedges) and observed no changes in the top 30 patterns. A few new patterns in the top 100 included a fourth argument (e.g., pattern #40, #50, #77, and below). We believe these patterns are not meaningful, particularly due to their low ranking, and therefore implemented the size constraint on hyperedges.

As mentioned in Chapter 3, type inference is performed during the generalization procedure with respect to the rules in Table 2.3. Consequently, inner parts of subedges may disappear, such as modifiers and triggers. There is the option in our system to expand subedges to their next level by setting the `expand` argument in `get_levels()`

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

to `True`. We tested our system with three different settings for the evaluation presented in Chapter 5: no expanded subedges, with additional expanded subedges (denoted with **all**) as well as exclusively expanded subedges (denoted with **exp**). Figure 3.5 shows two patterns derived through expanded subedges.

Further key configurations are available for the evaluation of learned patterns. Most importantly, the user can determine the number of patterns to be considered for the extraction process. We tested our system with the top 1, 5, 7, 10, 15, 20, 25 and 30 patterns. Additionally, it is possible to vary the maximum number of distinct extractions allowed per decomposed hyperedge. In our experiments, the number ranged from 1 to 5, with 3 being used most frequently. Lastly, if the user of our system focuses on extracting binary relations, additional arguments from the `LSOIE` gold tuples can be ignored for the evaluation. This scenario is denoted with **bi** for "binary" in the next chapter.

4.3 `WiRe57` Scorer

This section summarizes the functionalities of the `WiRe57` scorer, which we use to evaluate our novel OIE system (see Chapter 5 for results). The scorer was introduced by Léchelle, Gotti, and Langlais [LGL19] alongside the `WiRe57` dataset, forming a fine-grained evaluation framework with high-quality annotations that enables a detailed assessment of an OIE system's performance. They stated that "matching a system's output to a reference is not trivial". Therefore, they focused on certain aspects such as manually crafted annotations (including implicit relations and coreference information) and partial matching at the token level. Their goal was to establish a reliable benchmark for fair comparisons between different OIE systems.

The evaluation relies on three metrics: precision, recall, and F_1 score, which are commonly used in information extraction tasks to measure the alignment between predicted and gold-standard extractions. In this context, precision measures the correctness of the extractions, while recall measures their completeness. The F_1 score is the harmonic mean of precision and recall, providing a balanced measure of overall performance. A detailed explanation of the matching and scoring procedure follows below.

The `WiRe57` scorer supports exact and partial matching. An extraction is considered an exact match if it exactly matches the gold tuple for each tuple component. This is a very strict scoring method and not suitable for testing our approach. Partial matching allows for more flexibility by scoring token-level overlaps between the predicted and gold-standard extractions. This is particularly useful in OIE tasks, where the same information can be expressed in slightly different ways, which can lead to small mismatches. Since both the `WiRe57` and the `LSOIE` datasets are token-based, the use of token-level partial matching aligns well with the structure of our extractions. Below, we briefly explain how the scoring process works for each sentence:

Let $G_s = \{g_1, g_2, \ldots, g_N\}$ be the set of gold tuples and $T_s = \{t_1, t_2, \ldots, t_n\}$ the set of extractions for a given sentence s. The first step is to compute tuple-tuple matching

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

scores, meaning that each gold tuple $g_j, j \in 1, \dots N$ is compared with each extracted tuple $t_i, i \in 1, \dots n$. We refer to a tuple as $t = (t^{a_1}, t^r, t^{a_2}, \dots) = (t^{p_k}), k \in [1, k_n]$, where p_1 represents the subject, p_2 the relation, p_3 the first object, and so on. The elements within each tuple component are treated as bags of words, thereby disregarding the original order.

Each subject, relation and first object of a gold tuple is compared *separately* to its corresponding component $t^{p_k}, k \in \{1, 2, 3\}$ of the extraction t. The number of matching words is defined as the number of predicted words that also appear in the reference. The number of reference words includes all "real" words of the gold annotation, excluding inferred words and coreference information, to guarantee fair comparisons between the systems. The number of predicted words is the total count of words in t^{p_k}. A partial match is granted if at least one token aligns for each of these three core components. This restriction attempts to avoid rewarding random extractions. However, there remains room for improvement, as function words like "the" or "of" are also considered valid matches. The matching of further objects follows a similar approach. If the gold tuple contains only three components, any additional objects in the system's extraction (if present) are ignored, and the metrics remain unchanged. For additional objects in the gold tuple ($g^{p_k}, k > 3$), each tuple part g^{p_k} is compared to its corresponding component t^{p_k} in the prediction t. By enforcing positional matching, the scorer ensures that the extracted information aligns not only in content but also in the intended logical order.

For the final tuple-tuple score between a predicted tuple t_i and a gold tuple g_j, the numbers of matching words is summed over k, i.e. across all tuple components. Precision measures the proportion of correctly extracted words relative to the total number of predicted words. Recall reflects the proportion of "real" reference words correctly identified in the system's predictions. The mathematical definitions of precision and recall in this setting are given in Equations (4.1) and (4.2).

$$\text{precision}(t_i, g_j) = \frac{\sum_k |t_i^{p_k} \cap g_j^{p_k}|}{|t_i|} \tag{4.1}$$

$$\text{recall}(t_i, g_j) = \frac{\sum_k |t_i^{p_k} \cap g_j^{p_k}|}{|g_j|} \tag{4.2}$$

$$F_1 = \frac{2 \cdot p \cdot r}{p + r} \tag{4.3}$$

Léchelle, Gotti, and Langlais [LGL19] employ a greedy algorithm to map the predicted tuples with the references for each sentence. The scorer computes the F_1 score as defined in (4.3) for all possible combinations and selects the (t_i, g_j) pair with the highest F_1 score. Both tuples are then removed from the set of candidate tuples, and the process is repeated until no further matches are found. Afterwards, precision, recall, and F_1 score are calculated at the sentence level, before aggregating the results across the dataset. Ultimately, the overall performance of an OIE system is reflected in its token-weighted precision and recall across all tuples.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 4.3 illustrates the strictness of the scoring process by comparing the binary gold-standard extraction (Nelson Mandela; did; great things for his country) with the 3-ary extraction (Nelson Mandela; did; great things; for his country) obtained through the post-processing step presented in Table 4.2. Although both tuples contain exactly the same words, they differ in how these words are assigned to tuple components (i.e., positions). Since scoring is performed per component (i.e., per annotation label), it is impossible to achieve 100% recall for this extraction. This is because "for his country" is assigned to **arg3** in the predicted tuple, while it belongs to **arg2** in the gold-standard reference. As the gold-standard tuple contains no additional objects, any extracted information assigned to **arg3** is disregarded. The precision remains at 100% since all predicted words in the relevant tuple components match those in the gold tuple (5 matching words out of 5 predicted words). However, the reference tuple contains more words (8 in total), resulting in a recall of only $\frac{5}{8} = 62.5\%$.

Variable	Gold Tuple	Predicted Tuple	#Matches
arg1	Nelson Mandela	Nelson Mandela	2
rel	did	did	1
arg2	great things for his country	great things	2
arg3	–	for his country	-

Table 4.3: Example of partial matching of gold vs. predicted tuple.

The `WiRe57` scorer is a suitable evaluation tool for our work since it has already been used to benchmark a range of other OIE systems (as discussed in Chapter 2). Its token-level partial matching approach aligns well with the structure of the `LSOIE` annotations, which ensures a meaningful and accurate assessment of our system and penalizes excessively long extractions. However, Kolluru et al. [KAA+20] noted that due to the scorer's one-to-one mapping strategy, systems generating extractions that combine information from multiple gold extractions are unfairly penalized. This limitation represents a significant disadvantage for benchmarking our system. During the evaluation process, we observed that recall values greater than 1 were occasionally returned for tuple-tuple scores. This malfunction occurs when a reference word appears more than once in the corresponding predicted tuple component. Another issue is that inferred words, which are part of 54% of the `WiRe57` gold-standard tuples, are not considered in the matching procedure. This can lead to seemingly successful matches even when the extraction fails to convey the intended information. For instance, if the gold-standard relation is entirely inferred, any word in the predicted relation could count as a match. As a result, systems that generate a large number of extractions may appear to perform better simply because one of them is likely to match such a gold tuple.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.4 Comparison Systems

As discussed in Chapter 2, several OIE systems have been evaluated using WiRe57, which has become a widely accepted benchmark in the field. We assess extractions from all 57 sentences produced by seven OIE systems, as provided by Léchelle, Gotti, and Langlais [LGL19], and compare our system's performance against them as well as the results obtained by Menezes and Roth [MR21].

We also aimed to evaluate some of these models on the LSOIE datasets. However, this was particularly challenging since many OIE systems are not publicly available, poorly documented, or require a complex setup, which significantly limited the number of eligible options. Ultimately, we successfully extracted triplets for the LSOIE sentences using ClausIE [DCG13]. Therefore, we used a Python wrapper[1] since ClausIE was originally implemented in Java. We attempted to run the system without modifications, however, due to punctuation marks in the LSOIE sentences, such as apostrophes and commas as thousand separators in numbers, we had to add a `try/except` clause to handle these cases without causing execution errors. To align with the guidelines of our approach, we only considered extractions that contain at least one object. Additionally, we discarded tuples where at least one part of the extraction consisted of only one character, assuming these resulted from parsing errors. We tested the system on all three LSOIE datasets. Although no model training was required for rule generation, the extraction procedure was time-consuming, processing approximately 1,200 sentences per hour.

Moreover, we intended to compare our system to the original approach developed by Menezes and Roth [MR21] on the LSOIE datasets. However, this was not straightforward, as their system is not accessible online. After contacting the authors, we obtained their evaluation script for the OIE task and attempted to replicate their results on the WiRe57 dataset first. The next best option was to test their rules (presented in Table 3.4) within our own evaluation setup, omitting the post-processing step of combining extracted tuples. Despite these efforts, we were unable to reproduce the results reported in the paper for the WiRe57 dataset. The closest match to the published results was achieved using their evaluation script, which is independent of `newpotato` as well as the `graphbrain` framework. To stay as close as possible to the original approach, we used this script to evaluate the LSOIE datasets with the 13 rules derived by Menezes and Roth, assuming that their original system would have produced similar results.

[1] https://github.com/drwiner/ClausIEpy (Last accessed: April 17, 2025)

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

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

CHAPTER 5

Evaluation and Results

This chapter presents the results of different strategies for learning valid OIE patterns (as described in Chapter 3) and for performing tuple extraction (see Section 4.1), aiming to answer RQ1 and to establish a final solution. We evaluate the performance of these approaches on four datasets from two distinct domains: `Wikipedia` articles and scientific texts. As explained in Section 4.3, precision and recall serve as the key metrics, measuring the correctness and completeness of the extractions. We also report the F_1 score, which balances precision and recall, providing a single metric that reflects accuracy and coverage. Finally, we compare our system against existing OIE baselines to address RQ2 and RQ3.

As discussed in Section 4.4, we attempted to reproduce the results obtained by Menezes and Roth [MR21] on the `WiRe57` dataset using various settings. Table 5.1 presents the evaluation metrics reported in their paper alongside our results obtained with the `newpotato` framework (denoted as `np`) and the authors' evaluation script.

	Total Results			Matches			
System	**Prec**	**Rec**	F_1	**Prec**	**Rec**	**#**	**#Extr.**
1 Rule (Script)	.487	.086	.147	.71	.85	35	51
1 Rule (Paper)	.475	.184	.265	.69	.85	74	107
1 Rule (`np`)	.579	.061	.110	.70	.87	24	29
5 Rules (Script)	.475	.234	.314	.68	.86	93	134
5 Rules (Paper)	.416	.326	.365	.70	.93	120	201
5 Rules (`np`)	.633	.102	.176	.74	.85	41	48

Table 5.1: Results for different setups evaluating the `WiRe57` dataset, using one or five of Menezes and Roth's rules. "Script" refers to the evaluation metrics obtained with their evaluation script, "Paper" to their publication, and `np` to our OIE system.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Since the original publication reports the highest number of extracted tuples, it also achieves the highest overall recall, which boosts the F_1 score. In contrast, our system produces significantly fewer extractions, only about 25% of those reported in the paper when using five rules, but obtains the highest precision of 63.3% with five rules and 57.9% with one rule. The metrics generated with Menezes and Roth's evaluation script are comparable to those published, though consistently lower. These findings support our decision to use their script for evaluating the original semantic hypergraph-based OIE system on the `LSOIE` datasets.

We briefly outline the evaluation process and the reasoning behind our decisions. The primary goal was to find a balance between extracting more information and minimizing noise. An overview of all key configurations is given in Section 4.2. We started with the edge case configuration of one pattern and a maximum of only one extraction per decomposed hyperedge (short: 1/1) and increased these parameters step-by-step. While maximizing F_1 score was the main objective, we also aimed to identify configurations that would favor either high precision or high recall. After finding a solid configuration for the base pattern set, we extended the evaluation by including patterns derived from expanded subedges. We also explored alternative strategies, such as disabling the post-processing step that combines extracted triplets, to study its impact on the overall performance. The best configuration from each dataset served as the starting point for evaluating the next dataset, allowing us to refine our approach based on previous insights.

We present scatter plots with recall on the x-axis and precision on the y-axis. To visualize the different experimental setups, we use two colorbars to represent our main parameters:

- Number of patterns (inner circle)
- Maximum number of extractions per decomposed hyperedge (outer circle)

To distinguish between pattern sets and extraction methods, we apply text labels for the specific scenarios, as explained in Section 4.2. If no text label with **all** or **exp** is shown, the base pattern set (without expanded subedges) is used. To improve readability, redundant points are omitted. The results are sorted in ascending order by the number of patterns and the maximum number of extractions. This ensures that omitted points stem from higher parameter values, making them less relevant. The point achieving the highest F_1 score (excluding the **bi** scenario) is also highlighted.

5.1 `LSOIE-sci` Dataset

The filtered `LSOIE-sci/test` dataset is quite large, containing over 6,000 sentences and 9,901 extractions, as shown in Table 4.1. Although there exists a `dev` set as well, it is very small, and we chose not to include it in our evaluation, as its results would likely be unreliable.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

#Patterns	#Max. Extr.	Setting	#Extr.	#Matches	P	R	F_1
1	1	–	3,697	1,836	**.435**	.131	.201
5	1	–	7,374	3,550	.407	.269	.324
5	3	–	7,411	3,557	.404	.279	.330
10	1	all	8,450	3,831	.386	.294	.334
10	2	all	9,770	3,885	.344	.309	.326
10	3	–	8,135	3,889	.402	.313	.352
10	5	–	8,136	3,889	.402	.313	.352
15	2	–	8,299	3,969	.401	.319	.355
15	2	all	10,042	4,027	.345	.321	.332
15	3	–	8,309	3,969	.400	.319	**.355**
15	3	all	10,299	4,033	.337	.324	.330
15	3	bi	8,309	3,970	.404	.337	.368
15	3	exp	10,288	3,964	.334	.317	.325
15	3	exp/nc	21,274	4,420	.185	.349	.242
15	5	nc	18,174	4,464	.210	**.361**	.265
20	3	–	8,309	3,969	.400	.319	**.355**
20	3	all	10,463	4,044	.333	.324	.329
20	3	all/nc	21,553	4,535	.186	.359	.245
20	3	exp	11,768	4,037	.299	.324	.311
20	5	–	8,310	3,969	.400	.319	**.355**
25	3	all	10,794	4,080	.326	.328	.327

Table 5.2: Performance metrics for different extraction settings on LSOIE-sci/test sorted by number of patterns and maximum number of extractions in ascending order.

Figure 5.1 and Table 5.2 illustrate the results for the LSOIE-sci/test dataset. The best overall F_1 score of 0.355 was achieved with the configuration of 20 patterns and a maximum of 5 extractions per hyperedge. However, when rounding to three decimal places, the same F_1 score could be reached with 15 patterns and a maximum of 2 extractions. This indicates that increasing the number of patterns and extractions beyond this point does not necessarily improve performance but might increase computational cost. Therefore, the configuration 15/2 can be considered a more efficient alternative, balancing precision (0.401) and recall (0.319) effectively.

To optimize for precision, the configuration 1/1 yielded the highest value (0.435), but at the cost of very low recall (0.131). Conversely, the highest recall (0.361) was obtained with the **nc** strategy using 15 patterns and a maximum of 5 extractions, showing that skipping the post-processing step increases recall at a drastic expense of precision. We also evaluated the pattern sets **all** and **exp**, and observed similar recall values but lower precision. As expected, the **bi** evaluation returned the highest F_1 score (0.368). Since this setting focuses on binary relations, it reflects how the system would perform when the user prioritizes binary structures over complete extractions.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

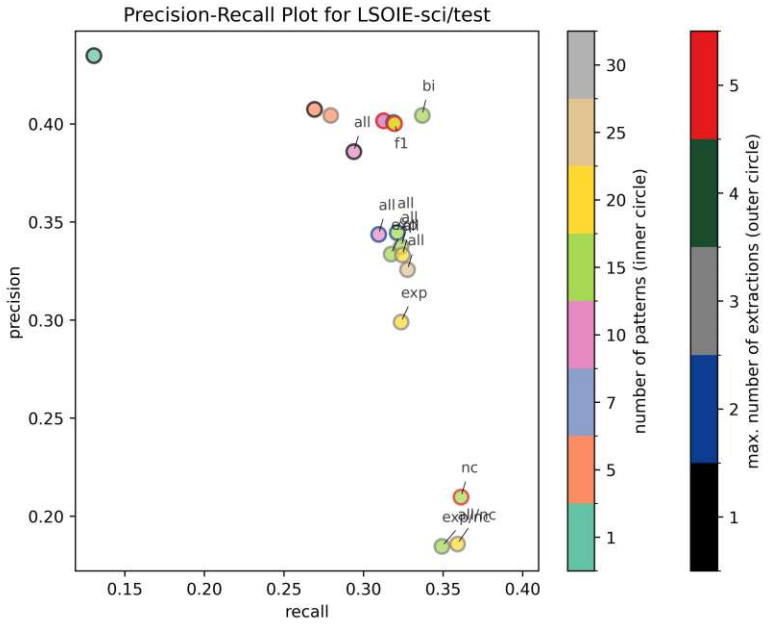


Figure 5.1: Results on `LSOIE-sci/test` for several key configurations.

System	LSOIE-sci/test P	R	F_1
Our system	.400	.319	.355
SH (5 rules)	.227	.182	.202
`ClausIE`	.065	.105	.080
`LSOIE` Baseline	–	–	.38

Table 5.3: Performance of all comparison systems on the `LSOIE-sci/test` dataset.

Table 5.3 presents the results on the reduced `LSOIE-sci/test` dataset for all available comparison systems. Our best configuration achieved a significantly higher F_1 score (0.355) compared to 0.202 for the original SH rules and 0.08 for `ClausIE`. However, the best neural baseline provided by `LSOIE` for the entire dataset reached an F_1 score of 0.38, which is slightly higher. Nevertheless, these results demonstrate that our rule-based system can compete with neural models.

5.2 `LSOIE-wiki` Datasets

Both `LSOIE-wiki` datasets, `dev` and `test`, contain nearly 3,000 gold tuples from approximately 1,500 sentences after data cleaning. Since the results for these two datasets are highly similar, the detailed analysis focuses on `LSOIE-wiki/test`. Table 5.4 and Figure 5.2 visualize our experiments for this dataset. The results of `LSOIE-wiki/dev`

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

are provided in the Appendix.

#Patterns	#Max. Extr.	Setting	#Extr.	#Matches	P	R	F_1
1	1	–	3,697	1,836	**.380**	.103	.162
5	3	–	2,246	984	.363	.248	.294
10	2	–	2,488	1,091	.361	.280	.315
10	3	–	2,495	1,091	.360	.281	.315
10	3	all	3,109	1,097	.300	.280	.290
10	3	exp	3,109	1,097	.300	.280	.290
10	4	–	2,495	1,091	.360	.281	.315
15	1	–	2,532	1,111	.363	.273	.311
15	2	–	2,545	1,112	.359	.285	**.318**
15	3	–	2,552	1,112	.358	.286	**.318**
15	3	all	3,201	1,135	.299	.291	.295
15	3	all/nc	6,196	1,329	.185	**.331**	.237
15	3	bi	2,552	1,113	.364	.304	.331
15	3	exp	3,191	1,115	.297	.285	.291
15	3	exp/nc	7,269	1,299	.157	.323	.211
15	3	nc	5,604	1,310	.197	.330	.247
15	4	–	2,552	1,112	.357	.286	**.318**
20	3	all	3,249	1,137	.296	.291	.293
20	5	–	2,553	1,112	.357	.286	**.318**

Table 5.4: Performance metrics for different extraction settings on LSOIE-wiki/test sorted by number of patterns and maximum number of extractions in ascending order.

The parameter combination 15/3 achieved the highest overall F_1 score (0.318), however, almost equal scores were obtained with a maximum of 2 or 4 extractions. The best recall (0.331) was reached with patterns including expanded subedges (**all**) and the **nc** scenario for the same parameters, but precision dropped significantly from 35.8% to 18.5%. In comparison to LSOIE-sci/test, similar results emerge for the edge case 1/1, the other pattern sets, and the **bi** setting.

	LSOIE-wiki/dev			LSOIE-wiki/test		
System	**P**	**R**	F_1	**P**	**R**	F_1
Our system	.358	.270	.308	.358	.286	.318
SH (5 rules)	.182	.166	.174	.174	.172	.173
ClausIE	.040	.072	.051	.039	.074	.051
LSOIE Baseline	–	–	–	–	–	.31

Table 5.5: Performance of all comparison systems on the LSOIE-wiki datasets.

Table 5.5 shows that our system achieved the highest F_1 scores on both LSOIE-wiki/dev

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

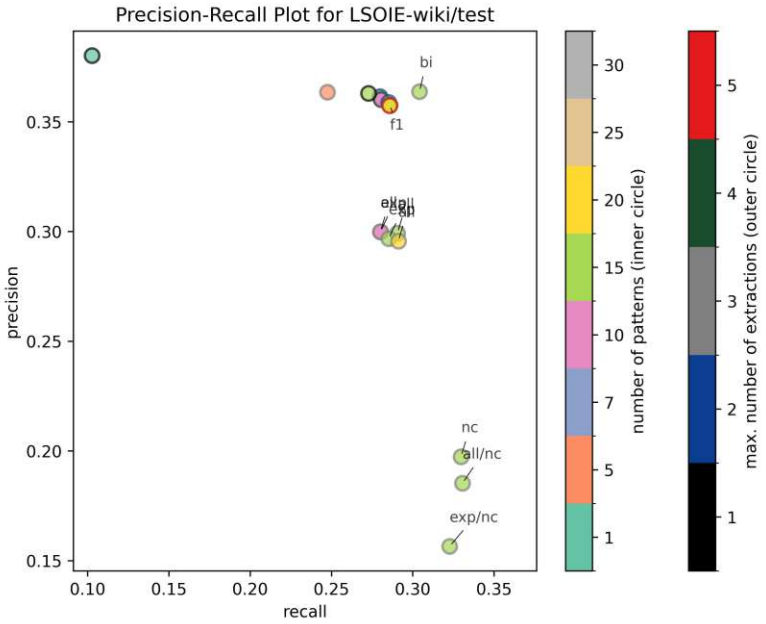


Figure 5.2: Results on `LSOIE-wiki/test` for several key configurations.

(0.308) and `LSOIE-wiki/test` (0.318), significantly outperforming the original SH rules and `ClausIE`. Furthermore, our system surpasses the F_1 score of the neural baseline provided by Solawetz and Larson [SL21] for the entire `test` set. A similar pattern is observed as for the `LSOIE-sci/test` dataset, though with slightly lower F_1 scores across all comparison systems.

5.3 `WiRe57` Dataset

This well-known dataset contains 343 tuples of 57 sentences. 57% of them are anaphoras and 54% include inferred words that are not part of the original sentence. Since most of the tuples (74%) represent binary relations, we omit the **bi** scenario. Figure 5.3 displays the tested configurations and the extensive list of experiments can be seen in Table 5.6.

The maximal F_1 score of 0.309 was achieved with the parameters 25/2 using the **exp/nc** setting. This result highlights that not combining similar triplets can lead to higher recall, which is particularly beneficial for this small dataset. We noticed that due to inferred words, which are disregarded for recall computation and tuple-matching, a lot of matches seem to be quite random. As a consequence, systems which produce a lot of noise, i.e. a high amount of similar triplets, can benefit from this limitation of the `WiRe57` dataset and its corresponding scorer. Once more, the edge case 1/1 achieved the highest precision of 56.9%, but recall amounts to only 5.3%. The recall (0.292) from our best configuration can be raised up to 0.338 by increasing the parameters to 30/5.

We observed in Figure 5.3 that the performance varied more strongly across configurations

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

#Patterns	#Max. Extr.	Setting	#Extr.	#Matches	P	R	F_1
1	1	–	34	26	**.569**	.053	.097
1	1	exp	34	26	**.569**	.053	.097
5	1	exp	73	51	.467	.107	.175
5	2	exp	78	52	.446	.118	.187
5	3	–	73	51	.470	.117	.187
10	1	–	79	55	.463	.120	.190
10	3	–	79	55	.467	.131	.205
10	3	exp/nc	229	109	.297	.244	.268
10	3	nc	168	100	.373	.224	.280
15	3	–	79	55	.467	.131	.205
15	3	exp/nc	238	110	.288	.246	.265
15	3	nc	184	105	.352	.238	.284
20	2	nc	164	101	.384	.218	.278
20	3	nc	185	106	.355	.240	.287
20	4	all/nc	262	114	.272	.257	.264
20	4	exp/nc	311	142	.283	.323	.302
20	5	–	79	55	.466	.132	.205
20	5	all/nc	276	115	.260	.260	.260
20	5	exp/nc	329	143	.269	.326	.295
20	5	nc	209	109	.323	.251	.282
25	2	exp/nc	246	132	.328	.292	**.309**
25	3	exp/nc	290	140	.298	.319	**.309**
25	4	exp	117	72	.409	.171	.241
25	4	exp/nc	320	145	.284	.332	.306
25	5	exp/nc	340	146	.268	.336	.298
25	5	nc	226	113	.311	.258	.282
30	2	all/nc	216	108	.312	.230	.265
30	3	nc	195	110	.352	.247	.290
30	4	exp/nc	335	146	.274	.334	.301
30	5	exp	129	72	.371	.174	.237
30	5	exp/nc	355	147	.259	**.338**	.293

Table 5.6: Performance metrics for different extraction settings on WiRe57 sorted by number of patterns and maximum number of extractions in ascending order.

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

than in the other test sets, likely due to different annotation styles. Surprisingly, the **exp** patterns outperformed the base patterns, reinforcing the assumption that diverse annotators may have influenced the results.

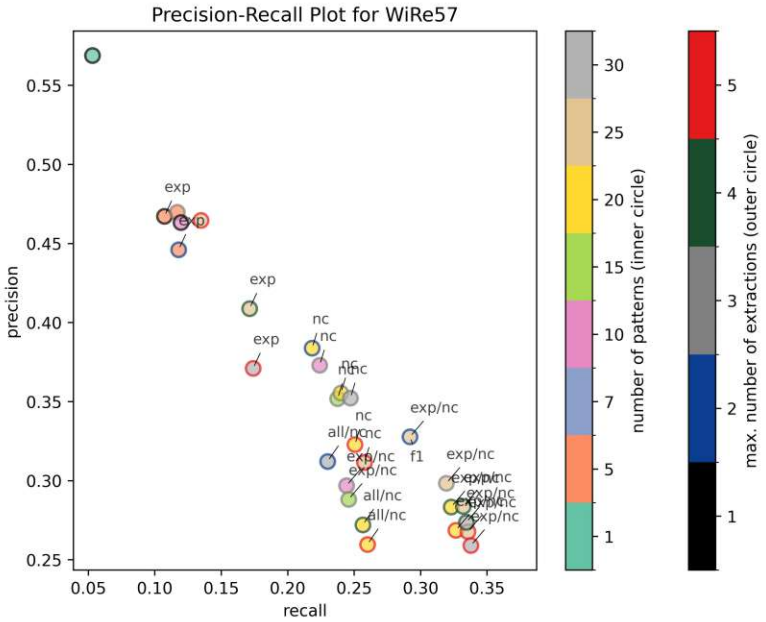


Figure 5.3: Results on WiRe57 for several key configurations.

	Total Results			Matches			
System	**Prec**	**Rec**	F_1	**Prec**	**Rec**	**#**	**#Extr.**
SH (5 rules)	.416	**.326**	**.365**	.70	**.93**	120	201
MinIE	.400	.323	.358	.75	.83	**134**	252
ClausIE	.401	.298	.342	.74	.84	121	223
Our System	.328	.292	.309	.61	.76	132	246
OpenIE 4	.501	.182	.267	.68	.84	74	101
SH (1 rule)	.475	.184	.265	.69	.85	74	107
OLLIE	.347	.175	.239	.73	.81	74	145
ReVerb	**.569**	.121	.200	**.83**	.77	54	79
Stanford OIE	.210	.188	.198	.79	.65	99	371
PropS	.222	.162	.187	.59	.80	69	184

Table 5.7: Performance of OIE systems on WiRe57, ordered by descending F_1.

We compared our best results for the WiRe57 dataset with the results published in the SH paper [MR21]. This contrasts with our evaluation of the LSOIE datasets, where we ran the comparison systems ourselves. This may have significantly influenced the reported results and could explain the differences compared to LSOIE. For broader comparison

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

insights, we also included the results of additional OIE systems evaluated with WiRe57, as shown in Table 5.7.

The original SH system with 5 rules achieved the highest F_1 score (0.365), closely followed by MinIE (0.358) and ClausIE (0.342). While our system produced a lower F_1 score (0.309), it maintained a balanced performance in terms of precision and recall. Notably, the recall of our system does not differ significantly from the comparison systems, but precision is considerably lower. ReVerb achieved the highest precision (0.569) but at the cost of low recall (0.121) and a small number of extractions. Exactly the same precision value can be reached by our edge case configuration of one pattern and one extraction per decomposed hyperedge, but with reduced recall (0.053).

5.4 Final System

Our system demonstrated consistent and balanced performance across the four test datasets, achieving competitive F_1 scores in both domains. On the one hand, it notably outperformed the other rule-based OIE systems on the LSOIE datasets, which might be due to the fact that our patterns were obtained from LSOIE data. The F_1 scores for the two Wikipedia-based datasets were slightly lower than for the scientific-related datasets, but still significantly higher than the baseline models. Comparisons with the neural baselines provided by LSOIE show that we outperform a neural model on the scientific LSOIE dataset and achieve similarly strong results on the Wikipedia-related one. However, these comparisons may not be entirely fair, as we evaluated a reduced dataset, while LSOIE reports results for the complete dataset. Nevertheless, this highlights the strength of rule-based systems and demonstrates that it is not always necessary to rely on opaque neural networks when a fully transparent and interpretable system is available. The F_1 scores for the two Wikipedia-based datasets were slightly lower than for the scientific-related datasets, but still significantly higher than the baseline models. On the other hand, our system produced the lowest F_1 score (0.309) for the WiRe57 dataset compared to the results for SH and ClausIE found in the literature.

The best F_1 score for WiRe57 was obtained with the **exp/nc** strategy, which highlights that combining similar triplets is not always advantageous since it depends on the structure of the dataset and the objective of the extraction. By providing flexible parameters, the user of our OIE system can decide whether to focus on high precision for a small number of extractions or to extract a broader range of information and boost recall. The configuration of 15 patterns and a maximum of 3 extractions per hyperedge appears to offer the most balanced results. Moreover, the user is able to include more sophisticated patterns obtained by expanded subedges. For the LSOIE datasets, the base patterns worked best. The **bi** evaluation, which simplifies the problem to binary relations, offers insight into how our system would perform if additional objects were ignored.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

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

CHAPTER 6

Discussion

This section reflects on the findings presented in the previous chapter. We identify key errors based on a manual analysis of a small part of `LSOIE-wiki/dev`, discuss the limitations of our approach, and outline potential improvements for future work.

6.1 Error Analysis

We distinguish between errors that emerged during the rule learning phase and those that occurred in the matching process. For the latter, we analyzed the first 103 sentences of the filtered `LSOIE-wiki/dev` dataset, which contains 196 gold-standard tuples. The key configurations for this evaluation included the top 15 patterns without expanded subedges and a maximum of three extractions per decomposed hyperedge.

6.1.1 Evaluation Errors

From the 103 analyzed sentences, we extracted 159 tuples, of which 71 matched the gold standard, resulting in a precision of 75% and a recall of 77% for the matched cases. For four sentences, no extraction could be obtained. We manually classified the mismatches, assigning each to a single, most suitable error category when multiple causes were identified. This analysis revealed four main sources of error:

Hyperedge structure (30%): The most common issue was the failure to find a matching pattern for the relevant parts of the semantic hyperedge, often due to overly nested hyperedge structures.

Additional objects (25%): Another major source of error involved mismatches with additional objects. This included cases where the first object from the gold tuple (often a time designation) was missing in the prediction (while further objects would have been correct), the first and second object positions were swapped, or an additional object from the gold tuple was already included in the first object of the prediction (or vice versa).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Imprecise extractions (24%): Primarily, mismatches were caused by inaccurate relation extraction (9%). Minor issues included uninformative arguments, such as "who", where coreference resolution would have been required (4%), as well as problems with conjunction decomposition, overlong extractions, and rigid subject-relation combinations (each 3%). Although not affecting matching, auxiliary words in the relation extraction often lowered precision (15%).

Dataset limitations (21%): We excluded sentences without an object annotation to ensure proper triplet extraction. However, the `LSOIE` data still exhibits certain limitations, such as omitted negation, swapped subject and object roles, or annotations that miss natural language or fail to reflect the intended meaning of the sentence.

6.1.2 Training Errors

This section summarizes the errors encountered during the rule learning process. Almost all sentences from the `LSOIE-wiki/train` dataset were successfully parsed into semantic hyperedges. However, around 27% of the tuples were skipped due to the inability to find an exact match between the annotation spans and the hyperedges. As mentioned earlier, this is often not a limitation of our system but rather due to incomplete or flawed `LSOIE` annotations, as illustrated in Figure 3.2.

An additional 12% of the training tuples were discarded during the integration of functional patterns into hyperedges. These structural issues arise from token ambiguities (e.g., duplicated words in a sentence), nested hyperedge structures, or tokenization inconsistencies such as hyphenated words being split during parsing. For instance, in some cases, the subedge mapped to a specific annotation label cannot be solely isolated within a semantic hyperedge. Consider the sentence "The new elections are scheduled to take place on February 2 of next year." Its corresponding gold tuple (The new elections; scheduled; February 2 of next year; to take place) was successfully mapped to the subedges shown in Figure 6.1. The problematic mapping occurs for `ARG2`: the subedge `((to/M take/P.ox) place/C)` cannot technically be separated from `((to/M take/P.ox) place/C (on/T [...])`, since `place/C` serves as an atom and can only be extracted either by itself or together with both surrounding subedges.

Although approximately 40% of the gold tuples could not be used during the training process, the dataset's large overall size of 46,015 tuples still provides a sufficient amount of data for effective rule learning, as demonstrated by the results presented in Chapter 5.

6.2 Limitations

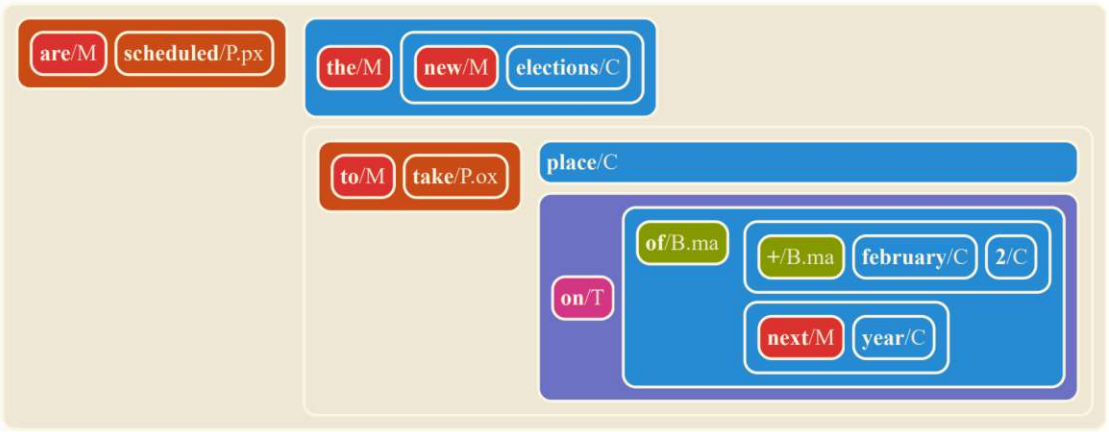
We encountered various challenges during our work. First, we had to reconstruct the OIE system using semantic hypergraphs solely based on the paper by Menezes and Roth [MR21], since it is not available in the `graphbrain` framework. When evaluating the original SH rules, we could not match the two patterns involving the special builder atom `+/B` due to a `graphbrain` malfunction in `match_pattern()`.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Sentence: The new elections are scheduled to take place on February 2 of next year.

LSOIE Annotation:
(The new elections; scheduled; February 2 of next year; to take place)

Semantic Hyperedge:



Tuple Mapping:

```
REL:   scheduled/P.px
ARG0:  (the/M (new/M elections/C))
ARG1:  (of/B.ma (+/B.ma february/C 2/C) (next/M year/C))
ARG2:  ((to/M take/P.ox) place/C)
```

Figure 6.1: Failed integration of functional pattern for ARG2 as its corresponding subedge cannot be solely extracted from the semantic hyperedge.

Another key issue was the inability to run other OIE systems because of undocumented and unmaintained GitHub repositories. Solawetz and Larson [SL21] provided a setup for executing state-of-the-art systems with their LSOIE datasets, as well as output files, but their instructions were outdated, and the systems' extractions were based on different test data (not the final LSOIE datasets used in our evaluation). As mentioned above, their annotations also showed inconsistencies.

The WiRe57 scorer also has limitations, noted at the end of Section 4.3, such as penalizing systems that combine information from multiple gold extractions, which is exactly the case for our OIE system. Furthermore, 54% of the tuples in the WiRe57 dataset contain inferred words, making it challenging to match with our extractions. The dataset's small size of only 57 sentences and 343 reference tuples may limit the statistical significance and generalizability of the results. A common issue for almost all OIE test datasets is that the ground truth cannot be assumed to be complete, which makes it difficult to measure recall.

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

6.3 Future Work

Several improvements could enhance the OIE system presented in this thesis. Expanding the training and test datasets to include a greater diversity of sources, such as `CaRB` [BAM19] or `BenchIE` [GYK+22], could improve the system's generalizability. Exploring alternative scoring approaches may provide a more balanced assessment of our system's performance. A broader comparison with other OIE models would offer deeper insights into its strengths and weaknesses. Moreover, architectural adjustments of `graphbrain` functionalities could reduce data loss during training and improve the accuracy of extracted patterns. If the `graphbrain` issue with the special builder atom `+/B` is resolved, patterns containing this atom could be used for extracting more diverse triplets. Finally, integrating coreference resolution could further enhance the quality of the extracted information.

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

CHAPTER 7

Conclusion

To address RQ1, we explored several strategies to find accurate patterns for OIE. We tested diverse pattern structures, including both basic and expanded hyperedges, to determine the effect of more sophisticated patterns on extraction quality. We also experimented with limiting the number of patterns and extractions per hyperedge to minimize noise and balance precision and recall. Furthermore, we evaluated the impact of post-processing extracted triplets by comparing scenarios with and without combining similar extractions. Initially, we adopted an unsupervised learning strategy but eventually switched to a supervised learning approach, which improved efficiency and eliminated manual effort. Ultimately, the combination of structural variation, parameter tuning, and supervised learning allowed us to develop a flexible and adaptable OIE system.

To answer RQ2, we compared the performance of our system against existing rule-based OIE solutions, specifically the original SH rules and `ClausIE`. Our system consistently outperformed both baselines on the `LSOIE` datasets. This may be partly due to a bias from our training data or an influence from the evaluation setup of the comparison systems. For the `WiRe57` dataset, our system performed worse than these baselines. However, it achieved better results than other state-of-the-art methods and maintained a balanced trade-off between precision and recall, highlighting the competitiveness of our approach.

Regarding RQ3, the generalizability of our learned patterns across datasets and domains appears promising. The stable performance for both scientific and `Wikipedia`-based datasets indicates that the patterns capture meaningful information, with slightly better results in the scientific domain. However, there remains significant potential for further testing on more diverse datasets, particularly those created by different annotators, to provide a more comprehensive assessment of the robustness of the symbolic patterns.

The main contribution of this thesis is the development of a fully transparent, rule-based OIE system that addresses key limitations of neural approaches. Unlike black box models,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

which often lack interpretability and reliability, our system provides clear insight into how extractions are generated. The symbolic patterns derived from supervised rule learning enable consistent and balanced performance across diverse datasets and domains. Our system allows users to adapt it to different extraction tasks while maintaining full transparency and efficiency. This work demonstrates that effective and competitive OIE solutions can be achieved without relying on complex neural architectures, emphasizing the value of symbolic approaches in NLP.

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

Overview of Generative AI Tools Used

As declared in the statement of originality, the use of generative AI tools for this thesis was limited to an auxiliary role, with my creative influence predominating. In particular, I have used `ChatGPT`, specifically `GPT-4-turbo`, as linguistic aid for my written work to rephrase my self-composed sentences in a more formal way which improves flow and clarity. Additionally, `ChatGPT` was used as a search engine, for instance, to find relevant literature.

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

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

List of Figures

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

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

List of Tables

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

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

Bibliography

[AJPM15] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. Leveraging linguistic structure for open domain information extraction. In Chengqing Zong and Michael Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China, July 2015. Association for Computational Linguistics.

[BAM19] Sangnie Bhardwaj, Samarth Aggarwal, and Mausam. CaRB: A crowdsourced benchmark for open IE. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 6262–6267, Hong Kong, China, November 2019. Association for Computational Linguistics.

[BGMMS21] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery.

[BLM+22] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What does it mean for a language model to preserve privacy?, 2022.

[CWZ18] Lei Cui, Furu Wei, and Ming Zhou. Neural open information extraction. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 407–413, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[DCG13] Luciano Del Corro and Rainer Gemulla. Clausie: clause-based open information extraction. In *Proceedings of the 22nd International Conference on*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

World Wide Web, WWW '13, page 355–366, New York, NY, USA, 2013. Association for Computing Machinery.

[FMHZ18] Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. Large-scale QA-SRL parsing. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[FSE11] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 1535–1545, USA, 2011. Association for Computational Linguistics.

[GGdC17] Kiril Gashteovski, Rainer Gemulla, and Luciano del Corro. MinIE: Minimizing facts in open information extraction. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2630–2640, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[GYK+22] Kiril Gashteovski, Mingying Yu, Bhushan Kotnis, Carolin Lawrence, Mathias Niepert, and Goran Glavaš. BenchIE: A framework for multi-faceted fact-based open information extraction evaluation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4472–4490, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[HLZ15] Luheng He, Mike Lewis, and Luke Zettlemoyer. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[KAA+20] Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. Openie6: Iterative grid labeling and coordination analysis for open information extraction, 2020.

[KAR+20] Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Mausam, and Soumen Chakrabarti. IMoJIE: Iterative memory-based joint open information extraction. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5871–5886, Online, July 2020. Association for Computational Linguistics.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[KGIR22] Ádám Kovács, Kinga Gémes, Eszter Iklódi, and Gábor Recski. Potato: explainable information extraction framework. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22. ACM, October 2022.

[KGR+22] Bhushan Kotnis, Kiril Gashteovski, Daniel Oñoro Rubio, Vanesa Rodriguez-Tembras, Ammar Shaker, Makoto Takamoto, Mathias Niepert, and Carolin Lawrence. milie: Modular & iterative multilingual open information extraction, 2022.

[LGL19] William Lechelle, Fabrizio Gotti, and Phillippe Langlais. WiRe57 : A fine-grained benchmark for open information extraction. In Annemarie Friedrich, Deniz Zeyrek, and Jet Hoek, editors, *Proceedings of the 13th Linguistic Annotation Workshop*, pages 6–15, Florence, Italy, August 2019. Association for Computational Linguistics.

[Mau16] Mausam. Open information extraction systems and downstream applications. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 4074–4077. AAAI Press, 2016.

[MR21] Telmo Menezes and Camille Roth. Semantic hypergraphs, 2021.

[MSS+12] Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. Open language learning for information extraction. In Jun'ichi Tsujii, James Henderson, and Marius Paşca, editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[SD16] Gabriel Stanovsky and Ido Dagan. Creating a large benchmark for open information extraction. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2300–2305, Austin, Texas, November 2016. Association for Computational Linguistics.

[SFDG16] Gabriel Stanovsky, Jessica Ficler, Ido Dagan, and Yoav Goldberg. Getting more out of syntax with props, 2016.

[SL21] Jacob Solawetz and Stefan Larson. LSOIE: A large-scale dataset for supervised open information extraction. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2595–2600, Online, April 2021. Association for Computational Linguistics.

[SMZD18] Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. Supervised open information extraction. In Marilyn Walker, Heng Ji, and

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[YBB+07] Alexander Yates, Michele Banko, Matthew Broadhead, Michael Cafarella, Oren Etzioni, and Stephen Soderland. TextRunner: Open information extraction on the web. In Bob Carpenter, Amanda Stent, and Jason D. Williams, editors, *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 25–26, Rochester, New York, USA, April 2007. Association for Computational Linguistics.

[ZZ20] Junlang Zhan and Hai Zhao. Span model for open information extraction on accurate corpus. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:9523–9530, April 2020.

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

Appendix

This appendix presents the results for the `LSOIE-wiki/dev` dataset, which are not discussed in Chapter 5. Table 1 summarizes the outcomes of all experiments conducted with different settings for this dataset. A corresponding visualization is shown in Figure 1.

#Patterns	#Max. Extr.	Setting	#Extr.	#Matches	P	R	F_1
1	1	–	953	424	.372	.093	.149
5	3	–	2,085	926	.354	.230	.279
7	2	exp	2,589	941	.299	.232	.261
10	2	–	2,300	1,029	.357	.261	.302
10	2	exp	2,768	1,018	.302	.256	.278
10	3	–	2,301	1,029	.356	.262	.302
10	3	all	2,846	1,018	.295	.257	.275
10	3	exp	2,846	1,018	.295	.257	.275
15	1	–	2,336	1,055	.362	.255	.299
15	2	–	2,345	1,056	.358	.269	.307
15	3	–	2,346	1,056	.358	.270	.308
15	3	all	2,928	1,071	.299	.273	.285
15	3	bi	2,346	1,056	.364	.291	.323
15	3	exp	2,926	1,044	.294	.264	.278
15	3	nc	5,623	1,270	.186	.316	.234
15	4	–	2,346	1,056	.358	.270	.308
15	5	–	2,346	1,056	.358	.270	.308
20	1	–	2,336	1,055	.362	.255	.299
20	3	all	2,973	1,071	.295	.273	.283
20	3	exp	3,281	1,073	.270	.272	.271
20	5	–	2,346	1,056	.358	.270	.308

Table 1: Performance metrics for different extraction settings on `LSOIE-wiki/dev` sorted by number of patterns and maximum number of extractions in ascending order.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

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

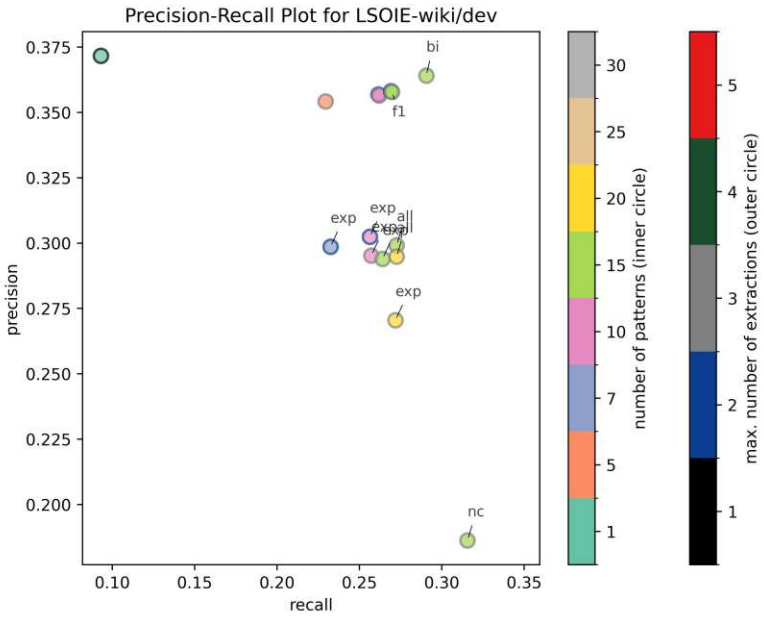


Figure 1: Results on `LSOIE-wiki/dev` for several key configurations.