



FAKULTÄT FÜR **INFORMATIK**

Patent Claim Decomposition for Improved Information Extraction

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Software Engineering & Internet Computing

eingereicht von

Peter Parapatics

Matrikelnummer 0225859

an der

Fakultät für Informatik der Technischen Universität Wien

Betreuung:

Betreuer: ao. Univ.-Prof. DI Dr. Andreas Rauber

Mitwirkung: DI Dr. Michael Dittenbach

Wien, 08.12.2009

(Unterschrift Verfasser)

(Unterschrift Betreuer)

Eidesstattliche Erklärung

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

Wien, 08.12.2009

Acknowledgements

Writing this thesis would not have been possible without the help of several people. First of all I would like to thank my parents, Helmut and Cecilia Parapatics, for their continuous support during my studies. Special thanks go to my supervisor Michael Dittenbach who has guided me through my thesis by spending numerous hours on discussions, contributing ideas and providing valuable feedback. I would also like to thank him for giving me the opportunity and working together with me on writing a scientific paper based on this thesis and submitting it to the CIKM 2009. I also want to express my gratitude to my supervisor Andreas Rauber, who helped me in choosing the topic of my thesis, for his ideas and comments. Thanks to Linda Andersson for reading the thesis and for providing additional ideas and feedback. Last but not least I want to thank Katharina Prochazka for proofreading and especially for her support while writing the thesis.

The patent documents analyzed in this work were kindly provided by Matrixware.

Zusammenfassung

Die vorliegende Arbeit beschreibt eine regelbasierte Methode zur Zerlegung von englischsprachigen Patentansprüchen in kleinere Teile mit dem Ziel, eine Basis für weitere Textanalyseschritte zu schaffen und die Anwendbarkeit von existierenden Algorithmen zur Informationsextraktion zu vereinfachen, welche auf Grund des komplizierten sprachlichen Aufbaus von Patentansprüchen nur beschränkt für diese geeignet sind. Da Patentansprüche nach sehr genauen syntaktischen und semantischen Vorgaben verfasst werden müssen, enthalten sie eine Reihe von wiederkehrenden grammatikalischen Mustern, die mittels linguistischer Analyse gefunden und extrahiert werden können. Die extrahierten Teile werden in eine Baumstruktur gebracht und es wird ein Algorithmus vorgestellt, der diese Teile reorganisiert und graphisch darstellt, um die Lesbarkeit der Patentansprüche zu verbessern. Die Evaluierung der Methode zeigt, dass die Länge und Komplexität von Patentansprüchen durch die Anwendung der entwickelten Regeln stark reduziert werden kann und dass dadurch die Anwendbarkeit von existierenden Information Extraction Tools erleichtert wird.

Abstract

Natural language processing algorithms and information extraction methods have proven to be valuable tools supporting humans in structuring, aggregating and managing large amounts of information, available as text, in several domains. Patent claims, although subject to a number of rigid constraints and therefore pressed into foreseeable structures, are written in a very domain-specific and almost artificial language common information extraction and retrieval methods tend to show poor performance on. This work presents a rule-based approach for decomposing patent claims into smaller parts for providing a basis for further analysis. As claims are drafted according to very precise syntactic and semantic rules, they contain a high number of reoccurring grammatical patterns. A set of rules based on linguistic analysis is used to identify and extract these patterns. The extracted claim parts are organized in a tree structure in order to retain the information on how they are related to each other. An algorithm is proposed for automatically reorganizing and then visualizing this tree structure for improving readability of claims. The evaluation of the method shows that rule-based patent claim decomposition is feasible and provides promising results in terms of reduction of length and complexity of patent claims. It shows that the decomposition method can be used to ease the application and raise the performance of existing information extraction tools.

Contents

1	Motivation	1
1.1	Introduction	1
1.2	Patent Search Problem	1
1.3	Goal of Work	2
1.4	Outline	3
2	The Patent Domain	4
2.1	Introduction	4
2.2	Intellectual Property	5
2.3	Patents	6
2.3.1	European Patents	6
2.4	Patentability	7
2.4.1	Novelty	8
2.4.2	Inventive Step	8
2.4.3	Industrial Application	9
2.4.4	Patentability Exceptions	9
2.5	Patent Documents	10
2.5.1	Document Types	10
2.5.2	Patent Structure	11
2.6	Patent Claim Structure	18
2.6.1	Independent Claims	18
2.6.2	Claim Categories	20
2.6.3	Dependent Claims	23
2.6.4	Claim Order	24
2.6.5	Terminology and Grammatical Structures	26
2.7	Patent Search	27
2.7.1	Types of Patent Search	27
2.7.2	Search Process	29
2.7.3	State of the Art in Patent Search	30

3	Related Work	34
3.1	Introduction	34
3.2	Structure analysis	35
3.3	Document Retrieval	35
3.3.1	Keyword-Based Approaches	35
3.3.2	Conceptual Search	38
3.4	Patent Similarity Measures	40
3.5	Parsing	40
4	Method	42
4.1	Data Set	42
4.2	Method Outline	44
4.3	Technologies	46
4.3.1	GATE	46
4.4	Regular Expressions	51
4.5	Document Parsing	52
4.6	Data Cleaning	55
4.7	Claim Type Identification	58
4.7.1	Dependent or Independent Claim Identification	58
4.7.2	Claim Category Identification	58
4.8	Claim Decomposition	60
4.8.1	Pattern Identification	60
4.8.2	Pattern Extraction	60
4.9	Independent Claim Decomposition	62
4.9.1	General Patterns	62
4.9.2	Physical Entity Claims	63
4.9.3	Method Claims	70
4.9.4	Use Claims	72
4.10	Dependent Claim Analysis and Decomposition	73
4.10.1	Reference Analysis	74
4.10.2	Claim Splitting	76
4.10.3	Refinement-Part Decomposition	78
4.11	Merging of Dependent and Independent Claims	78
4.11.1	Merge Process	79
5	Evaluation	86
5.1	Independent Claims: Length and Complexity Reduction	86
5.1.1	Length Reduction	86
5.1.2	Complexity Reduction	87
5.2	Quality Estimation of Independent Claim Decomposition	89
5.2.1	Physical Entity Claims	89

5.2.2	Method Claims	102
5.3	Claim Merging	105
6	Application	109
6.1	Claim Tree Visualization	109
7	Conclusions and Future Work	111

Chapter 1

Motivation

1.1 Introduction

The number of patent applications filed per year is continuously increasing. Over 50 millions of patents exist worldwide [TLB⁺09] with an increasing number of patent filings each year. According to the World Intellectual Property Organization (WIPO)^{1,2} about 1.75 million patents were filed in 2006 representing an increase of 4.9% compared to 2005. Because of their large economic value many industries ranging from pharmaceutical to information and communication technologies show major interest in patents. In addition to that it is claimed in scientific literature [Sch00] and by commercial patent retrieval service providers³ that a large amount of technological knowledge, between 80% and 90%, is only contained in patent documents, making them a valuable source of information.

1.2 Patent Search Problem

Although the quality of results obtained by patent searches has significantly improved with the emergence of new search engines and databases, there is an increasing concern that relevant documents may be missed [Atk08]. According to [KR09] 6% of patents are incorrectly rejected and 10% are incorrectly granted. The reasons are on the one hand the increasing number of patent applications being filed and on the other hand the lack of retrieval tools especially tailored to patent documents. Since most tools and processes

¹http://www.wipo.int/ipstats/en/statistics/patents/wipo_pub_931.html

²All Web links referred to in this thesis have been checked for existence and validity on 2009-12-08.

³<http://www.svpg.de/en/Patent/Default.asp>

are derived from tools used in other information retrieval fields they can not handle certain peculiarities of patent documents. Normal vocabulary is often used differently than in everyday language and grammatical structures which would be unthinkable in common texts are used routinely. The importance of having available good information retrieval tools is underlined in the Guidelines for Examination in the European Patent Office (EPO Guidelines) where it is stated that: “[...] it must be realised that in a search of this kind, 100% completeness cannot always be obtained, because of such factors as the inevitable imperfections of any information retrieval system and its implementation [...]” [EPO, Part B, III–1].

This is especially true for patent claims, which are usually very long and complex sentences, because of the requirements on the claim structure (described in detail in Section 2.6). Many automatic analysis tools tend to show poor performance on them. For example natural language processing (NLP) tools which are trained on general language texts like news paper articles have difficulties with the domain-specific vocabulary used for describing complex inventions. Additionally, certain keywords, which are used in patent claims with a different than their common meaning, lead to inaccurate parsing results for entire sentences. An example is the keyword “said” which is used in patent claims for referring to an already introduced concept and not as a verb. Often inventors intentionally use non-standard vocabulary in order to prevent search system from finding relevant prior art documents [Lar99].

Unlike in other domains it occurs frequently that an element of an invention and modifiers to it can be separated by a considerable amount of text. This is mainly a result of the use of independent claims for introducing new elements and dependent claims for refining them. Therefore proximity operators which work well in other domains may perform poorly in the patent domain [Atk08].

The very domain-specific vocabulary and the complex grammatical structures which are used in drafting patent claims do not only effect the performance of information retrieval tools but make claims very difficult to read for non patent experts.

1.3 Goal of Work

The claims in a patent can be seen as its essence, because they legally define the scope of the invention while the description and drawings have a supporting role to make the patent document easier to understand. The goal of this work is to develop a decomposition method for English-language patent claims. As claims are drafted according to very precise syntactic and

semantic rules a rule-based approach is chosen over a statistical method. It is investigated how the structure of the claim-specific language can be used to split the long and complex claim sentences into smaller components in order to improve the performance of natural language processing tools such as dependency parsers. The method identifies constituents and relations in claims, extracts them and puts them into a machine-processable structure for further analysis. For providing an application example the extracted parts are rearranged and merged into a tree structure which can then be visualized for improving the readability of claims.

1.4 Outline

The thesis is organized as follows. Chapter 2 introduces the patent domain to the reader by describing legal backgrounds, the structure of patent documents with a focus on syntactic and semantic particularities of claims and state of the art patent search techniques. In Chapter 3 an overview over related work in the field of patent information retrieval is provided. Chapter 4 describes the method developed in this work and gives a detailed explanation of the used decomposition rules. Chapter 5 provides an evaluation of the method and Chapter 6 shows how the extracted parts are visualized as a tree in order to improve readability of patent claims. In Chapter 7 the thesis is summarized and open issues and future work are discussed.

Chapter 2

The Patent Domain

Abstract

This chapter provides an overview of the patent domain. It starts with an introduction to the chapter in Section 2.1 followed by a brief description of the intellectual property domain in Section 2.2. In Section 2.3 a definition of what a patent is and which purpose it serves is provided. In order to provide a better understanding of the legal background Section 2.4 describes patentability requirements and exceptions to patentability. The main parts of this chapter are Section 2.5 in which the structure of a patent document is described and Section 2.6 focusing on syntactic and semantic particularities of the claim section. In Section 2.7 the four most important patent search types are described, the structured search approach used in the European Patent Office is introduced and state of the art patent search techniques are explained.

2.1 Introduction

It is important to introduce the patent domain to the reader before going into detail on state of the art patent information retrieval methods and describing the method developed in this work. Due to the fact that the method in this work is developed on and primarily for European patents the explanations are mostly based on the European Patent Convention (EPC) [EPC07] and its interpretation by the European Patent Office (EPO) in the Guidelines for Examination in the European Patent Office (EPO Guidelines) [EPO]. The Manual of Patent Examination Procedure (MPEP) [MPE08] used by the United States Patent and Trademark Office (USPTO) is used as a supplementary source of information, due to large similarities between European

patents and US patents regarding patentability requirements as well as rules for examining and thus drafting patents.

2.2 Intellectual Property

*Intellectual property (IP) refers to creations of the mind: inventions, literary and artistic works, and symbols, names, images, and designs used in commerce.*¹

This very broad informal definition found on the homepage of the World Intellectual Property Organization (WIPO) gives a good indication of the wideness of the intellectual property domain. The WIPO does not attempt to formally define what intellectual property is but provides a list of various subject-matters protected by intellectual property rights ranging from literary, artistic and scientific works over industrial designs to trademarks, service marks, commercial names and designations [WIP, page 3].

Intellectual property rights do not protect specific physical items but the knowledge and information reflected in these items. So considering a technical invention as an example, the produced item itself would not be protected but only the knowledge necessary for producing it.

The WIPO divides the intellectual property domain into two sub-branches: Copyright and Industrial Property.

Copyrights state that only the creator, or persons and organizations authorized by the creator, have the right to make copies of artistic work like books, audio compositions or motion pictures [WIP, page 4].

Industrial Property Besides patents, which are described in detail later in this chapter, the industrial property branch consists of the following fields [WIP, pages 8–15]:

- Utility models for protecting less complex technical inventions. Utility models are subject to reduced granting requirements compared to patents but also have a shorter term of protection.
- Industrial designs protecting an aesthetic aspect, reproducible by industrial means, of an industrial item like its form, material or color rather than the technical invention.

¹<http://www.wipo.int/about-ip/>

- Trademarks covering a sign or a combination of signs for distinguishing products or services of one company from those of a competitor.
- Geographical indications and appellations of origin which are associated with products of certain nature and quality like “Champagne” or “Tequila”.
- The layout and design of integrated circuits.
- Protection against unfair competition meaning any act of competition contrary to honest practices like creating confusion regarding the origin or manufacturer of a product.

2.3 Patents

The European Patent Office provides the following definition of a patent on their homepage:

*A patent is a legal title granting its holder the right to prevent third parties from commercially exploiting an invention without authorization.*²

The protection offered by a patent is not described in the patent document itself but regulated by the patent law of the country for which the patent is granted. A product patent usually gives the owner the exclusive right to “prevent third parties without the owner’s consent from making, using, offering for sale, selling or importing for these purposes the product” [WIP, page 7]. A process patent protects the use of the process itself and the use of products directly obtained by the process. The term of a patent is usually 20 years after which the invention enters the public domain and can be used freely. Therefore patents are a vital factor for assuring technical progress by protecting knowledge and as a result making research commercially interesting.

2.3.1 European Patents

With the EPC the currently 36 contracting states³ agree on a common law for granting patents which are called European patents [EPC07, Art. 2(1)]. European patents are subject to standardized patentability requirements,

²<http://www.epo.org/patents/Grant-procedure/About-patents.html>

³<http://www.epo.org/about-us/epo/member-states.html>

common regulations for the content of a patent application and a standardized examination procedure. One of the main advantages of a European patent lies in the fact that by filing a single application a patent which is valid in one or more contracting states can be obtained. Unless the EPC states otherwise it confers in each state the same rights as a national patent [EPC07, Art. 2(2)]. With the EPC the European Patent Office was established as a central authority responsible for granting European patents [EPC07, Art 4].

2.4 Patentability

According to the EPC a European patent shall be granted if the subject-matter of the patent refers to an invention from any technical field [EPC07, Art. 52(1)]. The subject-matter of a patent can be defined as the matter or the content for which protection is sought. In order to be patentable the EPC [EPC07, Art. 52(1)] requires an invention to

- *be new,*
- *involve an inventive step, and*
- *be susceptible of industrial application.*

Before describing these patentability requirements for inventions in more detail it is important to elaborate what the EPC considers an invention. The EPC does not define the term invention itself but provides an, according to the EPO Guidelines [EPO, Part C, IV–1], non-exhaustive list of things which are not considered an invention [EPC07, Art. 52(2)]. Most of the elements are not considered an invention due to the requirement implicitly contained in the EPC that “an invention must be of both a concrete and technical character” [EPO, Part C, IV–1]. Due to their abstractness mathematical theories, aesthetic creations and schemes, rules and methods for performing mental acts, playing games or doing business are not patentable under the regulations of the EPC. Mere discoveries, programs for computers and presentations are not considered an invention due to the lack of a technical effect. These exceptions apply only if the subject-matter of the patent directly refers to one of these exception. As an example: The mere discovery of a specific property of a material like resistance for heat is not patentable itself. But if this quality of the material is used for example to construct a heat shield, this heat shield would be considered a (possibly) patentable invention.

If the subject-matter of a patent application is an invention according to the EPC, it has to be assessed whether the invention fulfills the requirements stated above. In addition to these requirements explicitly stated in the EPC the invention has to fulfill the implicit requirement that it “must be such that it can be carried out by a person skilled in the art” [EPO, Part C, IV–1]. A person skilled in the art is described as an ordinary practitioner having a common general knowledge in the art [EPO, Part C, IV–22].

2.4.1 Novelty

An invention is only patentable if it “does not form part of the state of the art” [EPC07, Art. 54(1)]. In the EPC state of the art is defined as any relevant information “made available to the public by means of an oral or written description, by use or in any other way before the date of filing of the European patent application” [EPC07, Art. 54(2)]. It should be noted that this definition of state of the art is extremely wide with respect to how the information was published as well as to the geographical location and the language in which the information was published.

Basically any piece of information from which the subject-matter can be derived “directly and unambiguously” [EPO, Part C, IV–18] is enough to take away novelty from a claimed subject-matter. Also any feature implicitly derivable from such information by a person skilled in the art leads to unpatentability of the claimed subject-matter. The requirement of novelty is not taken away if the relevant information was under any bar of confidentiality or made available through an unauthorized person in an illegal act.

2.4.2 Inventive Step

The second required criteria is that the invention incorporates an inventive step. The criteria is fulfilled if the invention can not be considered obvious to a person skilled in the art with respect to the state of the art [EPC07, Art. 56]. According to the EPO Guidelines it is assumed that this person has had access to all material and knowledge in the state of the art [EPO, Part C, IV–22]. If the claimed subject-matter can be concluded directly from that information and does not go “beyond the normal progress of technology” [EPO, Part C, IV–23], the invention does not fulfill the requirement of incorporating an inventive step. Therefore a patent can not be granted for the invention. Nevertheless the EPO Guidelines point out that obviousness has to be evaluated with respect to the state of the art only. If for example

an invention is based on a new discovery which, once known, makes the invention trivial, the invention still fulfills the requirement of incorporating an inventive step [EPO, Part C, IV–23].

2.4.3 Industrial Application

The EPC considers an invention as susceptible of industrial application “if it can be made or used in any kind of industry, including agriculture” [EPC07, Art 57]. The EPO Guidelines suggest a very broad interpretation of the term industry. The main idea is to distinguish inventions having a technical background and being related to practical arts from any work in aesthetic arts. The requirement for industrial application does therefore not imply that an invention must incorporate a useful aspect. It rather requires that an invention is not an abstract concept but can be put into functionality. An example of an invention that can not be considered susceptible of industrial application is a perpetual motion machine which clearly contradicts well established physical laws and can therefore not be operational [EPO, Part C, IV–13].

2.4.4 Patentability Exceptions

The EPC excludes inventions from patentability for which a commercial exploitation would be contrary to “ordre public” or “morality” [EPC07, Art. 53]. A border case arises when an invention may have both an offensive and non-offensive use. Such inventions are patentable under the EPC as long as the subject-matter does not explicitly describe the use of or the construction of a machine for an offensive purpose [EPO, Part C, IV–7]. Due to the focus of this work on patents taken from the dentistry domain it should be noted that special rules are defined for the medical field where “methods for treatment of the human or animal body by surgery or therapy and diagnostic methods practiced on the human or animal body” [EPC07, Art 53(c)] are excluded from patentability. Patents may, however, be obtained for inventions like apparatuses, products or compositions used for these purposes. It should be noted that the EPC only excludes these methods from patentability if they are carried out on the living body [EPO, Part C, IV–11]. A patent for a dental implant can be considered as an example. The dental implant itself is patentable, while the process of implanting the device is not.

2.5 Patent Documents

2.5.1 Document Types

An identifier is assigned to each European patent document. The identifier is composed of a seven digit unique application number and the prefix “EP” indicating that the patent application was filed at the EPO. Additionally each document has a kind code assigned to it which is composed of a letter providing a coarse-grained status classification and a digit for refining the status. In this work the kind code of a referenced patent document is always shown after the identifier separated by a hyphen (“-”) resulting in a representation such as “EP1234567-B2”. A list of basic kind codes can be found on the homepage of the EPO⁴. The most important ones are described in the following paragraph.

Kind Codes Documents with the kind code “A” are patent application documents published after being filed with the EPO. “A” documents are further classified according to whether the application has already been published with a search report. The search report is a document created during the patent examination procedure citing all documents available to the EPO relevant for assessing whether the invention is novel and contains an inventive step. The following list of “A” documents can be found on the homepage of the EPO⁴:

- *A1 document: European patent application published with a search report*
- *A2 document: European patent application published without a search report*
- *A3 document: Separate publication of the European search report*

Documents with the suffix “B” are specifications of granted European patents. A European patent may be limited or even revoked by the owner of the patent, for example in order to strengthen the patent in view of some newly discovered prior art document, even after it has been granted [EPC07, Art. 105b]. This results in a change of the patent specification and is indicated by the digit following the letter “B”.

⁴<http://www.epo.org/patents/patent-information/european-patent-documents/basic-definitions.html>

The following list of “B” documents can be found on the homepage of the EPO⁵:

- *B1 document: European patent specification of a granted patent*
- *B2 document: New (amended) European patent specification*
- *B3 document: European patent specification after limitation procedure is complete*

2.5.2 Patent Structure

A European patent consists of several parts. Each part has a distinct purpose and is subject to specific requirements regarding its content and structure. The following parts will be described in detail in the sections below.

- Bibliographic Data
- Abstract
- Description
- Drawings
- Claims

Bibliographic Data

The bibliographic data section contains a large amount of metadata for a patent. It contains the title and the classification of the technical field of the invention. Furthermore it contains a number of important dates such as the application date, holds a list of designated contracting states, provides references to related documents which describe prior art and the background of the claimed subject-matter and contains information about the inventors and the patent applicants.

According to Rule 41 of the EPC the title of the invention shall “clearly and concisely state the technical designation of the invention and shall exclude all fancy names” [EPC07, Rule 41(2)(b)]. During the examination procedure of the EPO it is checked that the title is not misleading and correctly indicates the subject of the invention provided in the description and especially the claim section.

⁵<http://www.epo.org/patents/patent-information/european-patent-documents/basic-definitions.html>

Each patent is classified according to the International Patent Classification (IPC) scheme by assigning it to one or more categories. The following section will provide a brief introduction to the IPC scheme and other commonly used classification schemes.

International Patent Classification The IPC scheme was established with the intention to create an internationally uniform way for classifying patents [IPC09, page 1]. Work on the scheme was started after the establishment of the EPC in 1954 but the first official version was published only in 1968. The scheme was periodically revised to improve the system and to take into account the ongoing technical development. The IPC scheme was created with the primary goal to establish an effective search tool vital for examining the property of novelty and to ensure that an invention contains an inventive step [IPC09, page 1].

The IPC scheme is organized hierarchically with Sections on the highest level followed by Classes and Groups.

Sections The Section which is the highest level of the hierarchy gives a very broad and coarse-grained idea of the subject-matter of the invention [IPC09, page 3]. Sections are identified by capital letters ranging from A to H. For each Section a list of Subsections is defined [IPC09, page 3]. These Subsections have an informative purpose only and do not form part of the final IPC category identifier. They provide a description of a given Section and help understanding which areas are covered by that Section. Section A for example which is “HUMAN NECESSITIES” contains the Subsections “AGRICULTURE”, “FOODSTUFFS; TOBACCO”, “PERSONAL OR DOMESTIC ARTICLES” and “HEALTH; LIFE-SAVING; AMUSEMENT”.

Classes Each Section is subdivided into several Classes which form the second level of the hierarchy [IPC09, pages 3–4]. Each Class has a title and is identified by a two-digit number which is rendered unique only together with the Section letter. The Subclasses into which each Class is subdivided form the third level of the hierarchy and are identified by a capital letter attached to the Class code [IPC09, page 4].

The following example shows how the IPC scheme is applied by considering the IPC category A61C. The name of this IPC category is “DENTISTRY; APPARATUS OR METHODS FOR ORAL OR DENTAL HYGIENE” where “A” is the Section identifier for the Section HUMAN NECESSITIES which is further refined by the attachment of “61” to indicate the Class “MEDICAL OR VETERINARY SCIENCE; HYGIENE” and com-

pleted by the letter “C” for indicating the Subclass “DENTISTRY; APPARATUS OR METHODS FOR ORAL OR DENTAL HYGIENE”.

Groups In order to provide an even finer-grained classification each Subclass is further subdivided into Groups which themselves contain several Subgroups on various hierarchical levels. This means that not all Subgroups are direct children of the Main Group they belong to but can be children of other Subgroups in order to provide a further refinement of the classification [IPC09, page 4–5]. The Main Group symbol follows – separated by a white space – the Class symbol and is represented by a one to three digit number, a forward slash (“/”) and the digits 00. Each Main Group has a title assigned to it. In the dentistry domain for example the IPC classifier A61C 1/00 refers to the Main Group “Dental machines for boring or cutting”. The Group identifier is only meaningful together with the Section/Class/Subclass classification as described above.

The Subgroups are indicated by digits different to 0 following the forward slash. Each Subgroup’s title is preceded by one or more dots indicating the level of the Subgroup. The Main Group A61C 1/00 for example contains, among others, the Subgroup A61C 1/02 with the title “. characterised by the drive of the dental tool” and the second level Subgroup A61C 1/05 with the title “. . with turbine drive”. The title of the Subgroup is only meaningful when read in the context of its parent (Sub-)group. In many cases it is formulated to form a complete expression when attached to the title of the parent Group. The title of the Subgroup A61C 1/05 would therefore be read as “Dental machines for boring or cutting characterised by the drive of the dental tool with treadle or manual drive”.

Figure 2.1 shows a screenshot taken from the WIPO homepage⁶ showing the Group/Subgroup hierarchy of the IPC classifier A61C 1/00 explained above.

A61C 1/00	Dental machines for boring or cutting
A61C 1/02	· characterised by the drive of the dental tools
A61C 1/04	· · with treadle or manual drive
A61C 1/05	· · with turbine drive
A61C 1/06	· · with electric drive
A61C 1/07	· · with vibratory drive, e.g. ultrasonic

Figure 2.1: IPC Groups A61C

⁶<http://www.wipo.int/classifications/ipc/en/>

Core and Advanced Level With the last reform, which was started in 2000 and completed in 2005, the IPC scheme was divided into a core and an advanced level in order to provide individual support for the needs of different users. Different revision schemes were introduced for the core and advanced level. While the core level is only revised every three years the advanced level is revised continuously [IPC09, page 2].

Industrial property offices like the EPO are required to classify their documents at least according to the core level which includes information about at least the Main Group and for some technical fields Subgroups with only a small number of dots meaning that their hierarchical level is not very deep below their Main Group. The advanced level is intended for searching in larger patent collections since it provides a more detailed classification by specifying finer-grained Subgroups. The advanced level is usually compatible with the core level meaning that Classes and high-level Subclasses used in the advanced level are usually the same as in the core level. Nevertheless the two levels may differ due to the different revision intervals [IPC09, page 6].

Other Classification Schemes In addition to the IPC classification, European patents, usually have a classifier from the internal European Classification system (ECLA)⁷ assigned to them. The ECLA is an extension built on top of the IPC. With 135,600 Subdivisions it contains about twice as many Subdivisions as the IPC. The identification of the ECLA Subgroup is attached to the full IPC classification and is indicated by a letter optionally followed by a digit and another letter. So, for example, for the IPC identifier A61C 1/05, the ECLA defines the Subgroup A61C 1/05B with the title “Ducts for supplying driving or cooling fluid, e.g. air, water” and a further refinement of this Subgroup with the identifier A61C 1/05B1 and the title “through the working tool, e.g. hollow burr”. As in the IPC the titles are only meaningful when interpreted in the context of the parent Group.

The USPTO also uses their own classification system for US patents called United States Patent Classification system (USPC) which has a different structure than the IPC system. A detailed introduction to the USPC can be found in [Fal02]. The USPC is translated to the IPC by using static concordance lists with the consequence that the translated IPC category is not necessarily the most accurate one for a US patent.

⁷<http://ep.espacenet.com/help?topic=ecla>

Abstract

The abstract of a patent, which should have a length of no more than 150 words [EPC07, Rule 47(3)], serves the purpose of giving a concise summary of the disclosure of the invention provided in the description and the claims. It should indicate the title of the invention, describe the technical field it belongs to and should be drafted in a way which helps understanding the technical problem and the key aspects of the solution [EPC07, Rule 47(1)(2)]. Its main purpose is to provide an efficient way of searching the technical field and to help assessing whether the patent application itself should be read in detail [EPC07, Rule 47(5)]. It therefore does not have any legal effect on the application, neither with respect to disclosure nor with respect to the protected subject-matter.

Description

In order to be granted a patent the invention for which protection is sought has to be disclosed so that it can be carried out by a person skilled in the art. This person is assumed to be aware of common general knowledge and to have at his hand the means and capacities of necessary for fulfilling routine work. For determining what is considered common or general knowledge the EPO Guidelines refer to information contained in basic hand or textbooks on the field of the invention [EPO, Part C, II-2]. Although sufficient disclosure is assessed on the basis of the entire application except the abstract the description can be considered as the main source of information for this purpose. The EPC provides guidelines for the structure of the description section which should be followed “unless, owing to the nature of the invention, a different presentation would afford a better understanding or be more concise” [EPC07, Rule 42(2)].

The description should start with a specification of the technical field the invention pertains to [EPC07, Rule 42(1)(a)]. As shown in Example 2.1 no detailed specification is required. The technical field of the invention is described in a single sentence.

Example 2.1 EP0412246-B1

This invention relates to apparatus for irradiating photocurable dental materials with both light and heat and with the supply of heat being variable.

The indication of the technical field should be followed by an elaboration of background art useful for understanding the invention and should cite

relevant documents if possible [EPC07, Rule 42(1)(b)]. The most important part regarding the disclosure of the invention is the actual description of the invention often named “Summary of the Invention”. For sufficient disclosure the technical problem the invention deals with and the solution to it have to be described precisely enough to be well understood [EPC07, Rule 42(1)(c)]. Additionally any advantageous effects in relation to prior art should be stated [EPO, Part C, II–4]. The EPO Guidelines do not insist on a description in a problem-and-solution form and therefore allow the patent applicant to choose the form that discloses the invention best.

Any drawings included in the patent application have to be briefly described [EPC07, Rule 42(1)(d)] which is usually done in a single sentence in a manner such as shown in Example 2.2.

Example 2.2 EP0412246-B1

Figure 1 is a longitudinal sectional view of a hand-held heat-and-light-transmitting assembly [...]; Figure 2 is a front view of the heat- and light source of Figure 1 taken along the lines 2-2 of Figure 1.

The EPC requires the description to contain “at least one way of carrying out the invention using examples where appropriate” [EPC07, Rule 42(e)]. A description of all essential features for carrying out the invention and putting it into practice is required for a sufficient disclosure [EPO, Part C, II–4]. Ancillary elements or steps well known or obvious to a person skilled in the art on the other hand should not be contained in the description in order to keep its complexity as low as possible. For a complex invention or for claims covering a broad field one example may not be enough for disclosing the invention [EPO, Part C, II–4]. If it is not clear from the description or the nature of the invention, it has to be explicitly stated how the invention is industrially applicable [EPC07, Rule 42(f)]. Clear requirements are provided for the terminology and symbols used in the description [EPO, Part C, II–6]. Technical terms should only be used if they are recognized and well understood in the field to which the invention belongs and should not be used with a different than their established meaning if this is likely to cause confusion. The same applies to units describing a physical property of an invention where the applicant is required to use the units recognized in international practice. An example is the obligatory use of the metric system. In general those symbols, signs and terms should be employed which are generally understood in the domain of the invention. Also trademarks and proper names are only allowed if it is clear which product they describe. The use of trademarks only indicating the origin or the manufacturer of a product must not be used unless the product is specified explicitly. The used terminology

is not only required to be used consistently within the description section but throughout the whole application [EPC07, Rule 49(11)].

Drawings

Due to the fact that drawings are not allowed to be included directly in the abstract, description or claim section they have to be provided in a separate section and are then referenced where required [EPC07, Rule 49(9)]. They are subject to various requirements regarding their size, color and quality [EPC07, Rule 46]. For example no text shall be included except of a few keywords where they are indispensable for understanding the drawing [EPC07, Rule 46(2)(j)]. Tables are usually not considered drawings and are included directly in the description section and are also allowed to be used in claims [EPO, Part C, III-3]. References to figures made in the description and claims section are enclosed in parentheses and can consist of numbers as well as letters having a form such as: “(1), (2a)”.

Claims

*The claims shall define the matter for which protection is sought.
They shall be clear and concise and be supported by the description.*

[EPC07, Art. 84]

The claims in a patent can be seen as its essence, because they legally define the scope of the invention and the extent of protection conferred by the patent. The description and the drawings are only relevant for disclosing the invention and have a supporting role to make the claim section easier to understand and to interpret [EPO, Part C, III-1]. The EPO Guidelines put emphasis on conciseness and clarity of claims. They must be drafted “in terms of the technical features of the invention” [EPO, Part C, III-1] and should therefore not contain any non-technical matters like for example statements describing commercial advantages. Terms used in the claim should have a well defined meaning clear to a person skilled in the art from the wording of the claim alone without having to consult the description section [EPO, Part C, III-5]. Patent claims are therefore drafted according to very precise syntactic and semantic rules which are described in the following Section 2.6.

2.6 Patent Claim Structure

The rules for drafting patent claims are partly stated in the country's patent law but mostly defined implicitly by the patent examination guidelines used in the responsible patent office. In these guidelines it is described how certain keywords and grammatical structures should be interpreted. These rules for examining and thus also the rules for drafting patent claims are quite similar internationally but there are variations from patent office to patent office.

2.6.1 Independent Claims

The claim section of a European patent application contains at least one independent claim describing the essential features of the invention. Each of these independent claims may be followed by one or more dependent claims describing or refining particular features of the invention [EPO, Part C, III–4]. Dependent claims are described in detail in Section 2.6.3.

Form Each independent patent claim is defined in a single sentence. It should start with a part indicating the general technical class of the invention and describing already existing prior art knowledge. It describes the elements or steps of the invention that are conventional or known and for which no protection is sought. This so called “preamble” should only refer to relevant prior-art features, which are features necessary for the definition of the claimed subject-matter [EPO, Part C, III–1] [MPE08, 608.01(i)(e)(1)]. The part following the preamble is referred to as the “characterizing portion” specifying those technical features for which, in combination with the features stated in the preamble, protection is sought [EPC07, Rule 43(1)(b)]. For US patents this part is called “claim body”. The terms “characterizing portion” and “claim body” are used as synonyms in this work. In the form suggested by the EPC the preamble and the claim body should be connected with either the keyphrase “characterized in that” or “characterized by”. The USPTO refers to this connecting keyphrase as “transitional phrase” [MPE08, 2111.02] while EPO Guidelines consider it part of the characterizing portion referring to the whole structure of preamble-transitional phrase- claim body as “two-part form” [EPO, Part C, III–1]. The claim shown in Example 2.3 provides a simple example for the two-part form.

Example 2.3 EP1405609-A1

Preamble

Root canal reamer for handpieces, with two or three flutes

Transitional-Phrase

characterized in that

Body

it has from 2 to 7 turns

The preamble as well as the claim body are usually more complex than in the example shown above. This is shown in the independent claim in Example 2.4.

Example 2.4 EP1384449-B1

An apical foramen locator (10) comprising: a power circuit (30) operable to generate a stimulus voltage across two electrodes (12 and 18) and across a reference resistor (32) connected to one of the electrodes (18); an impedance-sensing circuit (34) operable to sense the stimulus voltage, a first voltage across the two electrodes (12 and 18) and a second voltage across the reference resistor (32);

characterized in that

at least one impedance map (72) including apical foramen location data (78) corresponding to a combination of a first voltage index and a second voltage index wherein the apical foramen location data (78) is generated from reference teeth; and a processing component (36) operable to derive the first and second voltage indices from the voltages sensed by the impedance-sensing circuit (34) and to select from the impedance map (72) apical foramen location data (78) that corresponds to the first and second voltage indices.

The whole part before the keywords “characterized in that” written in bold letters is considered prior art and only those features described in the claim body are protected by the patent.

The EPC states that the two-part form shall be applied “wherever appropriate” [EPC07, Rule 43(1)] like for inventions describing the improvement of distinct parts in a known combination [EPO, Part C, III–2]. For some inventions, however, the two-part form is considered inappropriate and is therefore not insisted upon. The EPO Guidelines provide the following examples which may require a different representation [EPO, Part C, III–2]:

- *the combination of known integers of equal status, the inventive step lying solely in the combination*
- *the modification of, as distinct from addition to, a known chemical process e.g. by omitting one substance or substituting one substance for another*
- *a complex system of functionally inter-related parts, the inventive step concerning changes in several of these or in their inter-relationships*

In practice it appears that a form different to the two-part form suggested by the EPO is used fairly often for drafting European patent claims. In most cases the alternative form is very similar to the drafting guidelines provided in the MPEP. As already stated above also the USPTO suggests the use of a preamble, a transitional phrase and a body. The difference lies in the keywords used in the transitional phrase where the most common ones are “wherein the invention comprises” or simply “comprising” [MPE08, 608.01(e)(2)]. The claim shown in Example 2.5 provides an example which is related to the third exception for the two-part form stated above.

Example 2.5 EP0154137-B1

A dental restoration comprising an outer shader layer (23), an intermediate layer (22) which is substantially hue and chroma free and translucent and an opaque substructure (21) which has a specific chroma on the Munsell scale and a specific Munsell hue.

If the two-part form is not used, it must be clear from the indication of prior art in the description which parts of the invention are prior art and which elements are the subject-matter for which protection is sought [EPO, Part C, III-3].

2.6.2 Claim Categories

Claims have different forms depending on which type of invention they describe. It can be differentiated between claims for physical entities and claims for activities [EPO, Part C, III-3]. Physical entity claims comprise claims for apparatuses and products while activity claims contain the subcategories “process or method claims” and “use claims”. As already stated in Section 2.4 abstract theories and mental acts are excluded from patentability, therefore activity claims always refer to processes or methods in which the use of some material is required [EPO, Part C, III-3].

Physical Entity Claims What is considered a product and what is considered an apparatus claim is not defined clearly in the EPO Guidelines. The category of physical entity claims contains claims for all types of products produced by a person's technical skill including substances and compositions like chemical compounds, entities like machines or apparatuses and any other physical object [EPO, Part C, III–3]. Example 2.6 shows a claim for a composition which starts with a short description of the type of composition and contains an enumeration of elements contained in the composition. Together with their descriptions these elements form the subject-matter for which protection is sought.

Example 2.6 EP1426413-A1

A polyorganosiloxane composition comprising (A) curable organopolysiloxane containing at least 5 mol% of diphenylsiloxane units or at least 10 mol% of methylphenylsiloxane units, curing agent for curing the organopolysiloxane (A), and (C) polyether having the compositional formula (1): $R_1O(C_2H_4O)_m(C_3H_6O)_nR_1$ [...].

Physical entity claims are normally drafted the following form: “An X, comprising a Y and a Z”. In most cases they include an enumeration of parts or elements an invention is composed of and a number of refinements to them. The placeholder X stands for the name or technical field of the invention while the placeholders Y and Z are replaced with the elements relevant for the protected subject-matter of the invention. This is illustrated by the claim shown in Example 2.7 in which the parts which indicate the field of the invention as well as the parts which describe elements of the invention and refinements to them are marked up.

Example 2.7 EP1405609-A1

Field of Invention	
A concentrated light source, [...] comprising:	
Element of Invention	
a bundle of fiber optic strands [...],	wherein the fiber optic strands [...]
	Refinement
Element of Invention	
a plurality of optical receptacles,	each receptacle optically coupled [...]
	Refinement

Method Claims Method claims have a very similar form to apparatus claims but instead of describing elements of a physical entity a sequence of steps is described. The form “A method for X comprising the steps of Y and Z” is very common where X is replaced with the goal of the method and Y and Z are replaced with a description of the steps necessary for achieving the goal. This structure is illustrated in Example 2.8.

Example 2.8 EP0154137-B1

Introduction
A method of preparing a dental restoration comprising the steps of:
Method Goal
Method Step
 preparing an opaque dental mount [...]
Method Step
 placing a crown which is [...] on said dental mount [...]

Use Claims Use claims are normally written in a form such as: “The use of X for Y” describing the use of a material or entity for achieving a certain goal. The place holder X is replaced with the material or entity which is used. The place holder Y describes the goal of the application of that material or entity. Example 2.9 shows a use claim for the manufacture of a chemical composition for cleaning teeth.

Example 2.9 EP0000256-B2

The use of a salt of lanthanum for the manufacture of a non-oxidising aqueous composition for cleaning plaque and/or stains from human teeth which consists essentially of the unbound cation of the element lanthanum in the form of a water-soluble salt and is free of any ingredients which precipitate the cation as a water-insoluble salt.

The interpretation of use claims is very similar to that of method claims [EPO, Part C, III–11]. The use claim above should therefore be regarded as equivalent to a process claim of the form: “A process of manufacturing a non-oxidising aqueous composition for cleaning plaque and/or stains from human teeth [...] using a salt of lanthanum”.

2.6.3 Dependent Claims

Dependent claims are used to refine elements of an invention and to introduce additional features. A dependent claim implicitly includes all features from all claims it references. It may only add further restrictions to the claim it refers to and is not allowed to broaden the claims. A dependent claim can refine more than one claim including other dependent claims. It can therefore refer to one or more independent claims, one or more dependent claims or to both dependent and independent claims [EPO, Part C, III–4]. It is important to mention that no cycling dependencies are allowed. A patent in which claim 1 depends on claim 2 and claim 2 contains a reference to claim 1 is not permissible. The obvious reason lies in the already stated requirement that a dependent claim must incorporate all features from the claim it refines and must always be more restrictive than the referenced claims.

Form The EPO suggests the following structure for dependent claims: The first part of the claim contains a reference to all claims it depends on followed by a second part which provides refinements of parts or a definition of additional elements of the invention [EPO, Part C, III–4]. The two-part form where the two parts are linked with “characterized in that” or “characterized by” is permitted and used frequently but is not required for dependent claims. The most common link phrases between the two parts is “wherein”. Occasionally the transitional phrase is not used at all and the refinement-part is started directly with a keyword like “comprising”. Two following examples should provide a better understanding of the structure of dependent claims. Example 2.10 shows the most common form of a dependent claim. It refers back to only one claim and adds a refinement to a part already defined in the referenced claim.

Example 2.10 EP0453689-A1

Reference
<div style="display: flex; align-items: center; justify-content: center;"> <div style="text-align: center; margin-right: 10px;"> { The bracket of Claim 1, </div> <div style="text-align: center;"> wherein Trans. Phrase </div> </div>
Refinement
<div style="text-align: center;"> { the tie wings are of substantially the same size. </div>

Example 2.11 shows a dependent claim which refines two claims. No transitional phrase is used and the refinement-part does not impose further

limitations on a specific element of the invention but adds a new element to it.

Example 2.11 EP0453689-A1

Reference

A dental implant anchor according to claim 1 or 2

Refinement

further comprising means internal of the anchor for engaging [...].

2.6.4 Claim Order

For European patent documents the EPO requires that the independent claims are grouped together with their dependent claims “in the most appropriate way possible” [EPC07, Rule 43(4)]. This leads to a structure where each independent claim precedes all its dependent claims. The dependent claims themselves are ordered from the least restrictive to the most restrictive. For US patents the USPTO defines the same order explicitly [MPE08, 608.01(n)].

Number of Independent Claims A patent has a clearly defined scope referred to as “the unity of invention” meaning that a European patent application “shall relate to one invention only or to a group of inventions so linked as to form a single general inventive concept” [EPC07, Art. 82]. The fact that a European patent may be granted not only for a single invention but also for a group of inventions together forming a single general inventive concept means that a patent may contain a plurality of independent claims. The EPO Guidelines interpret the EPC rule for unity of an invention by explicitly allowing the following combinations of independent claims [EPO, Part C, III–17]:

- *an independent claim for a product with an independent claim for a method or process for manufacturing the product and an independent claim for using the product*
- *an independent method claim with an independent claim for an apparatus for carrying out the process*
- *an independent claim for a product with an independent claim for a process or method for creating the product and an independent claim for an apparatus for carrying out the process*

The use of more than one independent claim of the same type in one patent application is not permitted. The only exceptions are the following inventive concepts described in [EPC07, Rule 43(2)]:

- *a plurality of interrelated products*
- *different uses of a product or apparatus*
- *alternative solutions to a particular problem, where it is inappropriate to cover these alternatives by a single claim*

The fact that more than one independent claim can be present leads to the fact that also independent claims may contain references to other claims without making them a dependent claim. The following two examples show two different ways of combining an apparatus claim and a method claim.

Example 2.12 EP0154137-B1

Independent apparatus claim:

A dental restoration comprising an outer shader layer, an intermediate layer which is substantially hue and chroma free and translucent and an opaque substructure which has a specific chroma on the Munsell scale and a specific Munsell hue.

Independent method claim:

A method for preparing a dental restoration by applying onto an opaque dental substrate having specific chroma on the Munsell chroma scale and a specific Munsell hue an intermediate layer, which is substantially hue and chroma free and translucent, and a shader layer.

Example 2.13 EP0474776-B1

Independent method claim:

A method of making stereophotographic documentation and photogrammetric measurement of impressions or models of jaws, comprising in combination the following moments: [...]

Independent apparatus claim:

An apparatus useable for carrying out the method according to claim 1 or 2, comprising in combination the following features: [...]

In Example 2.12 the independent method claim, which describes a method for creating the apparatus, does not contain a reference to the independent

apparatus claim but contains a copy of the description of the apparatus. In Example 2.13 the independent apparatus claim, which describes an apparatus for carrying out a method, contains an explicit reference to the independent method claim similar to the form used for references in dependent claims.

2.6.5 Terminology and Grammatical Structures

There are various grammatical structures which are commonly used in patent documents like enumerations of elements of an apparatus or steps of a method. Due to the fact that each claim has to be written in one sentence claim drafters use specific grammatical structures for chaining and nesting separate sentences to create a single sentence. Usually these nested sentences are used for providing a refined description of an already introduced part of an invention. There are various keywords or keyphrases for introducing such grammatical structures and for providing a specific semantic meaning for the claim. Two already introduced keyphrases are “characterized in that” and “characterized by” which are used for separating the preamble from the claim body.

Common keywords for introducing an enumeration are for example “comprising” or “consisting”. Although used synonymously in everyday language the keywords have a different semantic in the patent domain [EPO, Part C, III–13]. An enumeration introduced by the keyword “comprising” implies that the invention contains at least the specified parts. Enumerations introduced with the keyword “consisting of” are interpreted as including only the specified elements or steps and no others. They can for example be used for chemical composition with the meaning that the composition only contains the specified elements and no additional ones.

Unambiguous Wording The unambiguous use of terms in claims is reflected by the fact that new concepts are introduced with an indefinite article (“a” or “an”) or as noun phrases in plural. Subsequent uses of the same element are preceded by either the definite article “the” or by the word “said”. This is illustrated in Example 2.14 which shows an independent claim in which a new part is introduced and referred to later.

Example 2.14 EP0028529-B2

New-Concept		New-Concept	
A scaler tip	having	an operative end [...]	
Ref-Concept		New-Concept	
said operative end	terminating in	a curved free	

Optional Features If features are introduced with keywords like “preferably”, “for example”, “such as” or “more particularly”, they are regarded as entirely optional [EPO, Part C, III–8]. Relative terms like “thin”, “wide” or “strong” should not be used in claims, unless they have a well defined meaning in the domain like “high-frequency” for an amplifier [EPO, Part C, III–7].

2.7 Patent Search

Patent search differs from other information retrieval fields as there are various search types with different goals and thus different requirements. Also the search process itself differs greatly compared to other domains.

2.7.1 Types of Patent Search

The quality of search results is usually measured in terms of two basic parameters, recall and precision, which are defined as following:

$$\text{Recall} = \frac{\text{Number retrieved relevant documents}}{\text{Number existing relevant documents}}$$

$$\text{Precision} = \frac{\text{Number retrieved relevant documents}}{\text{Number retrieved documents}}$$

The recall value provides a measure for how many relevant documents a search was able to retrieve without taking into account how many non relevant results were contained in the results set. A high recall – a value close to 1 – would therefore mean that only a few relevant documents were missed by the search. The precision on the other hand indicates the accuracy of the search. A precision close to 1 means that the result set contains almost only relevant documents. It does, however, not provide any information about how many relevant documents have not been retrieved. There is usually an

inverse relationship between precision and recall which means that methods retrieving a high number of relevant documents (high recall) also retrieve a high number of non relevant documents (low precision) and the other way around. As a result search requirements can either target precision or recall. While for many information retrieval fields like ad-hoc Internet searches precision is the key factor, this is usually not the case for patent information retrieval, where missing a single document may have severe economic consequences. Therefore in most patent search tasks recall is more important than precision.

The four most common patent search types are informative searches, patentability searches, validity searches and infringement searches [Fog07].

Informative Searches

Informative searches – also called state-of-the-art searches [HNR07] – are usually conducted with the goal of getting an overview of what is currently being developed in a specific field of technology. Due to the fact that a significant portion of technical knowledge is documented only in patent documents [Sch00] a state-of-the-art search can provide macroscopic data for various purposes like research & development planning, competitor analysis or economic investment planning [Fog07]. For this type of search a high precision is usually more important than a high recall and the search effort can be kept relatively low compared to other patent search types.

Patentability Searches

A patentability search has the purpose of verifying that an invention fulfills the patentability requirements described in Section 2.4. It is usually conducted by the applicant prior to preparing and filing the application and by examiners in the patent office before a patent can be granted. The most important requirement that has to be assessed is novelty. Due to the fact that any document, available before the filing of the application, may impact novelty a patentability search should not miss a single piece of relevant information contained in patent and non patent documents. For patent information retrieval in particular this means that the full patent specification has to be analyzed [HNR07]. The very broad search field and high recall which is required render this type of search a very time consuming task.

Validity Searches

Validity searches are conducted with the goal to locate evidence that a patent claim was granted erroneously due to the oversight or concealment of prior

art information during the patent examination process [HNR07]. A common reason for a validity search is a potential patent infringement committed by a company for which it is sued as a result. If the company can prove that the patentability requirements for a claim were not fulfilled at the application's date of filing, the claim is invalidated and thus not enforceable. The search may also be conducted by the patent holder before going to court in order to avoid costly litigation. A validity search is also described as "a posteriori" [Fog07] or "exhaustive" [HNR07] patentability search.

Since the goal of this search is not the invalidation of entire patents but the invalidation of single claims each relevant claim of the patent has to be analyzed in detail rendering validity search a very time consuming task. Usually independent claims are the primary target since the invalidation of an independent claim results automatically in the invalidation of all its dependent claims. The search can usually be stopped when a satisfactory set of relevant documents is found and therefore does not need to reach a total recall [Fog07].

Infringement Searches

Infringement searches are performed in order to determine whether an enforceable patent claim exists which may have an influence on the exercise of a desired industrial activity, like the launch of a new product [Fog07][HNR07]. Infringement searches are therefore also called freedom-to-operate searches. Since the infringement of any in-force patent document can result in high financial losses no relevant documents must be missed by the search, thus a recall value of 1 is required [Fog07]. In addition to that a detailed analysis of the claim section for each retrieved document is required in order to assess whether relevant parts of a product are protected by the patent. The search can be restricted to patent documents which are still in force while already expired patents as well as non-patent literature can be excluded from the search.

2.7.2 Search Process

A professional patent search done by patent attorneys and patent examiners is a complex and structured process in contrast to ad-hoc queries as common for example in Web searches. Patent offices like the EPO or the USPTO suggest an iterative process and provide search guidelines for conducting a structured search [EPO, Part B, IV][MPE08, 904]. In [Fog07] and [Mic06] a basic approach for a structured search used at the EPO is described. It starts with the identification of relevant search concepts for each essential

feature of the invention. For each of these concepts keywords and classifiers (e.g.: IPC categories) are retrieved and documented in a search table. This is done by executing an initial search with only a few keywords describing each concept. From the result list the classification codes and common keywords are extracted using statistical methods. Those keywords and classification units which are likely to retrieve relevant documents are then combined in search queries. During the search process these parameters are iteratively adjusted and fine-tuned by using synonyms and adapting classification units based on the quality and the number of results. The search process should be stopped when evidence is found that patentability requirements can not be fulfilled or, according to the judgment of the examiner, “the probability of discovering further relevant prior art becomes very low in relation to the effort needed” [EPO, Part B, IV–6].

2.7.3 State of the Art in Patent Search

Patent Search Techniques

Keyword-Based Search In the keyword-based approach a patent search engine is used to retrieve patent documents which contain certain search terms in a specified section of the document. A basic keyword search can easily be conducted by an untrained user and is even likely to return relevant results [Sch00], but it can usually not be used for performing an exhaustive search.

The power of a keyword-based search engine largely depends on the expressiveness and flexibility of the query language. The more precise a query can be formulated the better the results are. In [Arc04] common query language features are analyzed. All state of the art search engines support Boolean operators like AND, OR and NOT for connecting search terms and allow the use of parentheses for nested expressions like (term₁ AND (term₂ OR term₃)). Normally search patterns can be specified by using wildcards for right-hand truncation like the asterisk (“*”) for matching words starting with a given sequence of letters (e.g. “document*” matches “document”, “documents”, “documenting”, ...). Some search engines also support the computationally more complex left-hand truncation by allowing wildcards at the beginning of a given sequence. For example the pattern “*oxide” would return all phrases ending with the sequence “oxide” such as “monoxide” or “hydroxide”. Besides left-hand and right-hand truncation, features like proximity operators, search term weighting and word stemming, provided for example by the patent search engine FreePatentsOnline⁸, are frequently

⁸<http://www.freepatentsonline.com/>

used in patent search.

Proximity operators allow the user to specify the maximum distance between two search terms. The expression “patent claim~5”, for example, means that no more than 5 words may occur between the two search terms for producing a match. Term weighting allows the searcher to specify the importance of a term. In the query “patent^5 or document” the word “patent” is five times more important to the relevance score of a document than the word “document”. Term weighting therefore influences the ranking of retrieved documents in the result list. Word stemming uses the word stem for the search meaning that a query for the word “scalars” also retrieves documents containing words like “scaler” or “scaled”.

Classification-Based Search In a pure classification-based search, all patents for a given category are retrieved and examined by hand. Due to the potentially large number of patents available for a given category this is a very time consuming task and thus often not feasible. A search with the esp@cenet⁹ search engine for the IPC third level Subgroup A61C13/03 (“Dental prostheses, with bases made of metal with a ceramic layer”) for example still returns 380 patents. In addition to that it is not always possible to find an exactly matching classifier for a given subject-matter.

Combined Approaches Usually a keyword-based search is combined with classification identifiers and additional non-subject information like application date, application number or publication date. Often a range can be provided as a search criteria for dates in order to, for example, retrieve all documents published between two given dates.

A keyword search can also be used to identify relevant classification categories for an invention. This is usually done by executing a keyword-based search and then using statistical means for finding categories which are assigned frequently to the retrieved patents [Sch00]. The patents in these groups can either be analyzed manually or the category identifiers can be used in combination with a refined keyword-based search. The WIPO provides a service called IPCCAT¹⁰ for determining likely IPC categories for keywords as well as for whole documents. The esp@cenet¹¹ search engine from the European Patent Office provides a similar service for the European Patent Classification system.

⁹<http://ep.espacenet.com/advancedSearch>

¹⁰<https://www3.wipo.int/ipccat/>

¹¹<http://v3.espacenet.com/eclsrch>

First Level vs. Value-Added Information

Patent information can roughly be divided into two categories: first level and value-added information. First level patent information is the original full text patent document. If the original patent information is enhanced and extended, it is referred to as value-added information. One of the first providers of value-added information was Derwent¹² which started as a value-added information database for pharmaceutical patents and was later extended to all areas of technology. Derwent provides rewritten abstracts and titles of patents for improving the findability of documents. In addition to that the database includes world wide documents with translated abstracts for improved retrieval of patents not written in English language. Other value-added patent databases are CAS REGISTRY¹³, a database for chemical research, or MARPAT¹⁴ which allows searching for chemical substances by their structure. Value-added information providers usually aim at integrating patents various sources for making their databases as comprehensive as possible.

While for a long time such databases have been the only providers of world wide patent information the patent search scene has significantly changed with the evolution of the Internet [Emm09]. Many providers of first level patent information like FreePatentsOnline¹⁵ or Google Patents¹⁶ have emerged. In addition to that major patent offices provide their own databases and search engines like esp@cenet¹⁷ by the EPO or the USPTO Patent Full Text and Image Database¹⁸.

With the emergence of first level patent information providers the question is coming up whether such – mostly free to use – databases can replace costly value-added information databases. It was shown that this is generally not the case and that free first level products expand the whole patent information market rather than making commercial value-added services redundant [Emm09, Sim06, Fog07, Phi05]. A case study in the pharmaceutical field [Emm09] has shown that a high number of relevant documents have only been retrieved from value-added databases. In addition to that the search precision was shown to be much higher than for first level documents for which the percentage of retrieved non-relevant documents was more than

¹²http://www.thomsonreuters.com/products_services/legal/legal_products/intellectual_property/DWPI

¹³<http://www.cas.org/expertise/cascontent/registry/index.html>

¹⁴<http://www.cas.org/expertise/cascontent/marpat.html>

¹⁵<http://www.freepatentsonline.com/>

¹⁶<http://www.google.com/patents>

¹⁷<http://ep.espacenet.com/>

¹⁸<http://patft.uspto.gov/>

30%. Nevertheless some documents were uniquely retrieved by a full text search leading to the conclusion that both, value-added as well as first level patent information, is needed for conducting a comprehensive search.

A detailed description was provided for the claim section for which the concepts of independent and dependent claims how they are related to each other was explained.

Summary

In this chapter the intellectual property and especially the patent domain were introduced. The term patent was defined, the purpose of a patent was examined and patentability requirements for an invention were analyzed. The focus of the chapter was the examination of the structure of a patent. For each part of a patent application – bibliographic data, abstract, description, drawings and claims – its purpose and structural particularities were examined. Different classification schemes for patents were introduced focusing on the International Patent Classification system. A detailed description of the claim section was provided. The concept of independent and dependent claims was explained and it was described how they are related to each other. The form used for drafting independent claims (preamble-transitional phrase-claim body) was described and various re-occurring grammatical structures were examined. Additionally three different classes of independent claims were introduced: physical entity claims, method claims and use claims. The structure of dependent claims was described and the semantic meaning of keywords and keyphrases commonly used in dependent and independent claims was introduced. Furthermore the four most important patent search types – informative searches, patentability searches, validity searches and infringement searches – were examined. The goal of each search type was analyzed and requirements in terms of precision and recall were described. Keyword-based, classification-based and combined approaches were introduced as state of the art search methods. Since a professional patent search is a complex task, the search process used by the EPO was examined. Additionally the contribution of first level and value-added patent information databases to patent search was analyzed.

Chapter 3

Related Work

Abstract

This chapter provides an overview of related work in the field of patent information retrieval and patent information extraction. In Section 3.2 a decomposition method developed for Japanese patent claims is described, which is very similar to the approach used in this work. Section 3.3 provides an overview over patent search and retrieval methods with a focus on associative retrieval methods, describing keyword-based approaches in Section 3.3.1 and conceptual search methods in Section 3.3.2. Section 3.4 describes a method for computing a similarity measure based on syntactic and semantic properties of patent claims and Section 3.5 describes a parsing method especially tailored to patent claims.

3.1 Introduction

Due to the large economic value of patents and the continuously increasing number of filed patent applications research is gaining importance in various fields of patent processing. A research focus is set on keyword and non-keyword-based patent document retrieval approaches. Various methods have been proposed for tackling the domain-specific challenges of patent information retrieval and patent information extraction described in Section 1.2. Other research fields are patent classification and summarization, readability improvement and patent structure analysis. Although several methods are developed for patents not written in English language (primarily Japanese documents) they are likely to be applicable also to English-language patents.

3.2 Structure analysis

In [SOMI03] an approach very similar to the method developed in this work is proposed for Japanese patent claims. Japanese patent claims are subject to similar structural requirements as European or US patent claims in the sense that they are written in one – usually long and complex – sentence which itself is composed of multiple sub-sentences. By manually analyzing Japanese patent claims the authors of [SOMI03] found six frequently occurring relations: Procedure, Component, Elaboration, Feature, Precondition and Composition. These relations are identified by keywords, so called cue phrases. The applied extraction algorithm consists of two steps. In the first step a lexical analyzer is used for differentiating between two types of tokens: cue phrase tokens and morpheme tokens. In a second step a parser generated from a handcrafted context free grammar is used for extracting the sub-sentences.

Patent claim decomposition can provide the basis for various purposes like machine translation, patent retrieval or improving readability of patent claims. The method described above for example displays the extracted sub-sentences in a graph structure for improving readability of claims. In [TFI04] for example the sub-sentences are used as a basis for an associative patent retrieval method.

3.3 Document Retrieval

A lot of research work is done in the field of patent search in order to improve and extend state of the art search methods as well as to develop advanced search methods especially tailored to the patent domain. Most approaches are associative document retrieval methods in which parts of a document or an entire document are used as query input for finding similar documents. Thereby the searcher is relieved from the difficult and time consuming task of finding relevant search terms. In the patent domain associative methods can be used for various search tasks like patentability or infringement searches and are especially useful for invalidity searches where documents similar to the claim section of one specific patent document have to be found.

3.3.1 Keyword-Based Approaches

In [TFI04] the authors exploit the fact that an invention and therefore the patent claims protecting the subject-matter of the invention contain more than one subtopic. Due to the requirement that a patent claim has to be

drafted in a single sentence more than one subtopic is usually contained in the same claim sentence. By extracting these subtopics and creating a single query for each of them the authors try to enhance search precision and therefore also the efficiency of the entire search process. In the first step of the method each compositional element of a claim (e.g.: each element of an invention) is extracted as a subtopic through the pattern matching method described in [SOMI03]. The second step of the method consists of extracting (mainly) noun and multi-noun sequences as query words from each subtopic. A predefined stop word list of 73 terms is used for filtering frequently occurring terms. In the third step an importance measure is calculated for each subtopic. The importance measure is calculated by summing up the specificity of each query term in the subtopic. The specificity value of a term is high if it appears only in a limited number of subtopics and low if it appears in many subtopics. In addition to that the importance value of a subtopic is higher if it is extracted from the claim body than if it is extracted from the preamble. In the retrieval process these values are used for computing a relevance score for each subtopic and each retrievable document in the text collection, based on the co-occurrence of query terms in the topic and the document. These sub-scores are summed up and weighted according to the importance measure computed for each subtopic in order to create the final relevance value for each retrieved document.

Retrieval Methods using Query Expansion

A similar approach is followed in [AF05]. The method is developed for retrieving patents written in English and Japanese taking a Japanese patent claim as query input. In a first step Japanese punctuation symbols are used as delimiters for segmenting the claim. Morphological analysis is performed for extracting nouns as query words. Too frequently occurring terms are filtered using a predefined stop word list. For searching English-language documents the extracted claim parts are automatically translated by a machine translation module. From the translated parts all words not contained in a stop word list are used as initial query terms. The biggest difference to the method described in [TFI04] is the so called “query expansion” step. From the top ten retrieved documents, the top ten terms, weighted according to a term frequency-inverse document frequency (TF-IDF) measure, are extracted for extending the set of initial query terms. In addition to that an intra document expansion method is applied. Each paragraph of the search input document is indexed. For each claim component matching paragraphs are retrieved based on the initially extracted keywords. The keywords from the matching paragraphs are then extracted and used for further extending

the set of initial query terms.

This extended set of keywords is then used to retrieve relevant documents for each claim component. The final relevance score for each retrieved document is computed from its weighted scores for each claim component.

Another two-stage retrieval method is proposed in [MMO⁺05]. The approach is developed for invalidity searches and takes, like the method described above, a patent claim as search input. In the first stage general text analysis and retrieval methods are applied to improve recall. In the second stage the search precision is improved by rearranging the top N documents. For this first stage query terms are extracted from the entire claim. Morphological analysis is performed and a hand-crafted stop word list containing 2,910 words is used for filtering frequently used words. A domain specific lexicon containing about 300,000 term entries is used for expanding the set of search keywords with semantically related terms. The search terms are weighted based on their IDF only without taking into account the TF. The extracted terms are then used for a keyword-based search. The target of this initial search is an entire patent document (abstract, description, claims). From the result set the top 1,000 documents with the highest relevance score are retained. Filtering methods, not described in detail by the authors, which are based on the IPC category, term subsets and passages are applied to filter out noisy documents. For the second stage the claims are divided into preamble and claim body by means of structural analysis. The preamble is ignored and only the claim body is used as query input. If a decomposition into preamble and body is not possible the entire claim is used. The pre-processing steps are the same as in the first stage but measurement terms like “speed” or “temperature” are assigned a higher weight. For identifying measurement terms a handcrafted dictionary with 361 entries is used. The weighted terms are used for executing a keyword-based search targeted at patent claims only. Other parts of a patent document like the description or the abstract are ignored. The relevance score from the second stage is added to the one from the first stage for producing a final relevance score for each retrieved document.

A single stage retrieval method using query expansion was already developed in 1999 [Lar99] for the USPTO for prior art searches. It allows the user to conduct natural language, Boolean and field searches and combinations thereof. The method uses an automated and a user guided query expansion component. The automated component adds phrases and compounds whose terms are already present in the original query. For this purpose a phrase and compound dictionary are used. If for example a search contains the words “tennis rackets” which can be found in the phrase dictionary, documents which contain these words in proximity are assigned a higher relevance

score. If the word sequence “tennis rackets” can be found as a single word (“tennisrackets”) in the compound dictionary it is added as a keyword to the query.

A wider range of possibly relevant phrases, which either contain the specified search phrase, or are associated with it are presented to the user and are only added to the query if the user selects them explicitly. The required co-occurrence data structure is computed offline and separately for each Class from the United States Patent Classification system. For this purpose the claims are divided into three sections: title and abstract, background information and claims. The co-occurrence measure is based on how often the phrase pairs co-occur in each section. The retrieval is done with Inquiry an information retrieval method based on Bayesian Networks and TF-IDF weighting [BCC93].

3.3.2 Conceptual Search

Due to the changing terminology and the frequent use of synonyms in patent documents the lexical matching of query words and documents has severe impacts on the efficiency of the search process. The searcher is required to have profound knowledge of the domain language in order to create queries by using every possible synonym for describing an invention. This problem is partly tackled by providers of value-added patent information databases like Derwent¹ by rewriting certain sections of a patent in a more standardized language. Nevertheless there is an increasing demand for conceptual search methods which are able to retrieve documents based on the semantic meaning of search terms.

The authors of [RSG08] motivate the use of Latent Semantic Indexing (LSI) [DDF⁺90] methods for performing a transition from lexical matching to semantic retrieval. Latent semantic indexing is a method in which an initial term-document matrix is transformed into a new, lower dimensional, coordinate system of conceptual topics. The recall and precision of LSI is highly dependent on the dimensionality reduction parameter “k”. In [MIW05] the use of LSI is evaluated on US patent documents and compared to a Vector Space Model (VSM) approach. Even with an appropriate value for “k” LSI was shown to perform only slightly better than a normal VSM. And for some values of “k” the recall and precision of LSI was even lower than with the VSM. The authors of [MIW05] therefore discourage the use of LSI for patent retrieval. In [RSG08] it is claimed that the poor performance might result

¹http://www.thomsonreuters.com/products_services/legal/legal_products/intellectual_property/DWPI

from inappropriately chosen preprocessing steps and parameters. They state that the performance of LSI could be improved by using a domain specific stop word list and a different stemming algorithm.

Large and heterogeneous document collections can pose problems for LSI. An inherent problem is that LSI can not handle polysemy of words. This means that the indexing algorithm can not differentiate between different meanings of the the same word. One out of many examples for this case is the word “dialog” which can be used in the computer domain for describing an input window rather than a conversation between two or more people. For reducing the heterogeneity and the size of a document collection the application of a clustering algorithm is suggested as a preprocessing step in order to break up a corpus into smaller homogeneous sub-collections [RSG08].

Another approach for challenging the inherent problems of the common LSI approach is proposed in [CTA03]. The authors suggest the use of an extended and more robust LSI approach, called Differential Latent Semantic Indexing (DLSI) in which the term-document matrix is replaced by interior and exterior differential document vectors. The interior differential document vectors are intra-document measures computed from different parts of a document or from the results of different summarization schemes applied to the same document. The exterior differential document vectors are computed from the document vectors of two different documents. Both matrices are transformed according to the LSI approach. For document retrieval both measures are combined for computing a similarity value. The DLSI method is used together with a template structure for storing patent document abstracts for associative document search. Abstracts are decomposed into parts and stored in a template structure rather than as full text. For a patent search the abstracts with the highest DLSI similarity measure are retrieved. A pattern matching approach is then used to refine the search. In an interactive process the template structure is enhanced with synonyms extracted from the query. This leads to a continuous improvement of the template structure and thus an improved search precision.

Structure-Based Conceptual Search

In [YS08] a similarity measure for patent claims is computed by comparing the structure of conceptual graphs extracted from the claims. The approach is based on the idea that if objects are represented as graphs, then similarity comparison of the graphs is equivalent to similarity comparison between the objects. A conceptual graph consists of relation and concept nodes. Concept nodes represent the entities of a domain while the relation nodes indicate how the entities are interconnected.

A conceptual graph G is defined as $G=(C,R,U,lab)$ with C being the concept vertices, R the relation vertices, U a set of edges for each relation and lab a set of labels. A label from the set lab is assigned to every vertex in the graph. A conceptual graph relies on the background support of a domain specific ontology for the concept and relation vertices. In [YS08] dependency relations generated with the Stanford Parser² are used for building the conceptual graphs. The graphs are then compared using a relaxation labeling method adapted for conceptual graphs. Relaxation labeling is a method, originally developed for image processing, for assigning appropriate labels to image objects taking into account the features of the objects' neighborhoods. It is an iterative process in which the support value for each label is continuously updated until all probabilities are stabilized. The authors adapt the method for finding the most likely matches between the conceptual graphs extracted from two patent claims.

3.4 Patent Similarity Measures

In [IAS07] a similarity score between two claims is computed based on a combination of simple lexical matching and knowledge-based semantic matching. A syntactic similarity score is computed based on the number of nouns that occur in both claims. The measure is proportional to the number of identical words and inverse proportional to the total number of words in order to avoid the score from being biased towards claims containing a larger number of words. A semantic similarity score is computed using WordNet [Fel98]. Each noun from the first claim is compared to all nouns from the second claim. The highest similarity score for each noun is recorded. The final semantic similarity score for two claims is calculated by summing up the semantic similarity score for each noun. The authors propose the use of the similarity score for categorizing patents and queries into topics in order to reduce the search space and increase the precision of a search. In addition to that the measure could also be used as a similarity score for an associative conceptual document retrieval method.

3.5 Parsing

To overcome the limitations of broad coverage statistical parsers like the Stanford Parser for the patent domain a complex domain specific parsing approach is proposed in [She03] and [Bab08]. The proposed parsing method

²<http://nlp.stanford.edu/software/lex-parser.shtml>

relies on supertagging [BJ99] and therefore on a domain specific shallow lexicon for annotating each lexeme with morphological, syntactic and semantic information. In [She03] the morphologic meta-data includes information about the part-of-speech tag (POS-tag) and the inflection type of words. An ontological concept defining the word membership to a certain semantic class (Object, Process,...) is used as semantic information. In the supertagging procedure each word is annotated with several matching supertags. In the following supertagging disambiguation procedure hand-crafted rules are used to eliminate contradicting supertags for each word by looking at a 5-word window to the left and to the right. The central part of the method is the predicate lexicon which is used to create a predicate-argument structure [SHWA03] by annotating each predicate with syntactic and semantic information. The semantic information includes a semantic class from the domain tuned ontology and several case roles. These case roles correspond to the arguments that have to be filled for the predicate. A lexicalized case role grammar is used to fill the required case roles for a predicate with chunked phrases, like noun phrases, from the claim based on the syntactic and semantic information in the supertag.

Summary

This chapter has provided an overview over existing research topics in the field of patent information retrieval and patent information extraction. A method developed for Japanese patents was described where claims are decomposed according to six frequently occurring relations identified by certain keywords. Since it also exploits reoccurring grammatical structures in claims it is very similar to the method developed in this work. Several patent retrieval methods were analyzed. The focus was set on associative document retrieval where query words are automatically extracted from certain sections of a patent to relief the searcher from the difficult task of finding appropriate query terms. Keyword-based search approaches in combination with query expansion methods were described. Additionally conceptual search methods which aim at retrieving documents based on the semantic meaning of search terms rather than on exact lexical matching were examined. An approach for computing a similarity measure between claims based on a syntactic and semantic similarity score was described. Finally a parsing method for patent claims based on supertagging was introduced. The method uses syntactic and semantic information to determine dependencies between constituents of a claim sentence.

Chapter 4

Method

Abstract

This chapter describes the decomposition method developed in this work. In Section 4.1 the data sets used for creating and evaluating the rules are described. Section 4.3 introduces the technologies and frameworks used in the method. In Section 4.4 the syntax of regular expressions used for describing the rules is introduced and commonly used symbols and abbreviations are explained. Section 4.2 provides an outline of the decomposition method which includes the steps of: extracting the claims from the original documents (Section 4.5), preprocessing them (Section 4.6), categorizing the claims (Section 4.7) and splitting them into smaller parts (Section 4.8). In Section 4.9 the decomposition rules for independent claims are described. Section 4.9.1 describes patterns extractable from all independent claim types. Sections 4.9.2, 4.9.3 and 4.9.4 describe the decomposition rules for physical entity claims, method claims and use claims. In Section 4.10 the decomposition of dependent claims is described. Section 4.11 describes how the tree structures extracted from dependent and independent claims are merged.

4.1 Data Set

For creating and evaluating the developed method two data sets from the IPC category A61C (Dentistry, Oral or Dental Hygiene) were used. A data set of 86 randomly selected patents was manually analyzed for creating the decomposition rules (Analyzed Set) and a larger set of 5,000 patents was used for evaluation (Evaluation Set). The Analyzed Set only consists of patents filed at the EPO while the Evaluation Set consists of 774 European patents and 4,226 US patents. The patents were sampled from the Matrixware Research

Collection (MAREC) data set¹. Tables 4.1 and 4.2 show the characteristics of the two data set. The first column indicates the claim type. The second column shows the number of independent and dependent claims in the data set. The third column shows the total number of words for both claim types and in the fourth column the average claim length can be found. The figures show that independent claims are more than three times as long as dependent claims.

	Nr. claims	Nr. words	Avg. claim length
Ind. claims	159	20,321	127.81
Dep. claims	862	28,794	33.40

Table 4.1: Analyzed Set: Characteristics

	Nr. claims	Nr. words	Avg. claim length
Ind. claims	13,628	1,803,341	132.33
Dep. claims	73,706	2,415,533	32.77

Table 4.2: Evaluation Set: Characteristics

The higher complexity of independent claims is also underlined by the high number of unsuccessful parses of independent claims as compared to dependent claims using the Stanford Parser. This is illustrated in Table 4.3 for the Analyzed Set and in Table 4.4 for the Evaluation Set. The first column indicates the analyzed claim type. The second column shows the settings for the maximum heap size for the Java Virtual Machine (JVM). The amount of memory available to the parser is an important parameter, because of the memory requirements for constructing the large parse trees for the relatively long independent claims. The fourth and fifth column show the number of successful and failed parses while the sixth column shows the percentage of successful parses.

It can be seen that the average number of successful parses is significantly higher for dependent claims than for independent claims. Additionally, the success rate of the parser decreases significantly when reducing the maximum heap size for the JVM. A successful parse in this context does not refer to the correctness of the parse tree but only indicates that the parser was able to produce a result. An informal evaluation of the parse trees indicates that the quality of the results is very low for the long and complex claim sentences. Example 4.1 shows a shortened claim and its parse tree which contains several

¹<http://matrixware.net>

Claim type	JVM max. heap size	Successful parses	Failed parses	% of successful parses
Ind. claims	1000MB	132	27	83.01%
	500MB	89	70	55.97%
Dep. claims	1000MB	859	3	99.65%
	500MB	848	14	98.38%

Table 4.3: Stanford Parser Success Rate: Analyzed Set

Claim type	JVM max. heap size	Successful parses	Failed parses	% of successful parses
Ind. claims	1000MB	10,671	2,957	78.30%
	500MB	7,482	6,146	54.90%
Dep. claims	1000MB	73,427	279	99.62%
	500MB	72,769	937	98.73%

Table 4.4: Stanford Parser Success Rate: Evaluation Set

errors. The noun phrase (NP) “a light emitting assembly within the housing” for example is attached to the verb phrase (VP) with the verb “having” rather than to the VP with the verb “comprising”. Additionally the word “means” is incorrectly tagged as a verb leading to an incorrect parse for the phrase “light transmitting means having a tip extending beyond the forward end of the housing”.

4.2 Method Outline

The method developed in this thesis consists of sequentially applied steps with the final goal of finding patterns which can be used to extract smaller sub-parts of a claim sentence. In the document parsing step, described in Section 4.5 the sections relevant for this method are extracted from the original patent documents. The transformed document is then parsed into an internal data structure using an XML parser. Before the extraction and splitting rules are applied simple preprocessing and data cleaning steps are carried out to remove unnecessary information and to normalize claims. Section 4.6 provides a detailed description of each preprocessing and data cleaning step applied to the original claim sentences. The main part of this method is the process of finding linguistic patterns which can be used to decompose the original claims. Due to structural differences individual extraction and splitting rules have to be applied to dependent and independent claims. For independent claims the patterns also vary depending on the category of the

Example 4.1 EP0219588-A1.xml: Incorrect Stanford Parse

An irradiation device comprising an elongate housing having a rearward end portion and a forward end portion, a light emitting assembly within the housing, light transmitting means having a tip extending beyond the forward end of the housing.

```
(ROOT
  (NP
    (NP (DT An)(NN irradiation)(NN device))
    (S
      (S
        (VP (VBG comprising)
          (NP
            (NP (DT an)(JJ elongate)(NN housing))
            (VP (VBG having)
              (NP
                (NP
                  (NP (DT a)(JJ rearward)(NN end)(NN portion))
                  (CC and)
                  (NP (DT a)(JJ forward)(NN end)(NN portion)))
                (, ,)
                (NP
                  (NP (DT a)(JJ light)(VBG emitting)(NN assembly))
                  (PP (IN within)
                    (NP (DT the)(NN housing))))))))))
          (, ,)
          (NP (JJ light)(NNS transmitting))
          (VP (VBZ means)
            (S
              (VP (VBG having)
                (S
                  (NP (DT a)(NN tip))
                  (VP (VBG extending)
                    (PP (IN beyond)
                      (NP
                        (NP (DT the)(JJ forward)(NN end))
                        (PP (IN of)
                          (NP (DT the)(NN housing))))))))))))
            )
          )
        )
      )
    )
  )
)
```

claim. In order to be able to apply these individual rules the claims have to be classified appropriately. For this purpose a set of heuristics is used which is described in Section 4.7. Although most decomposition rules are claim category specific, the pattern extraction process described in Section 4.8 remains the same. In Section 4.9 the extraction rules applied to independent claims are described. Section 4.10 describes how dependent claims are analyzed and decomposed. After all claims in a patent have been decomposed an algorithm is applied which tries to attach the refinements from dependent claims directly to the claim part where the refined element is introduced. This process is described in Section 4.11.

4.3 Technologies

The method is developed in Java SE6² and makes use of several open source frameworks for various tasks. For XML document parsing and manipulation JDOM³ is used as an easy to use alternative to the SAX and DOM parsers for Java. In the processing steps described in Section 4.11 a string similarity measure is required. For this purpose the open source Java library of similarity and distance metrics called SimMetrics⁴ is used.

Natural language processing which provides the basis for all extraction and splitting rules is done with the NLP framework GATE⁵ (General Architecture for Text Engineering).

4.3.1 GATE

Various NLP tools are available for Java. OpenNLP⁶ for example is a set of standalone tools which can be used for tasks like sentence detection, tokenization, part-of-speech tagging (POS-tagging), text chunking and dependency parsing. A more advanced framework is LingPipe⁷ which offers in addition to basic NLP modules features like named entity recognition, text classification and clustering. The main advantage of GATE compared to other tools is that it provides an integrated framework for a variety of NLP tasks. In addition to that the well defined API presents an easy way to integrate the framework into a Java NLP application.

²<http://java.sun.com/>

³<http://www.jdom.org/>

⁴<http://www.dcs.shef.ac.uk/~sam/simmetrics.html>

⁵<http://gate.ac.uk/>

⁶<http://opennlp.sourceforge.net/>

⁷<http://alias-i.com/lingpipe/>

Architecture of GATE

The architecture of GATE is based on the idea that elements in an NLP application can be broken down into various types of components called resources. Following the principles of component-based software development these resources are not hardwired but communicate with each other through well defined interfaces. Three categories of resources are differentiated in the framework architecture [Gat, Section 1.3].

- **Processing Resources** are algorithmic entities such as POS-taggers, natural language parsers and classification algorithms.
- **Language Resources** are entities which are processed or used by the processing resources such as text documents, text corpora, ontologies and dictionaries.
- **Visual Resources** are display and editing components used in graphical user interfaces.

GATE's architecture is based on plugins communicating with each other through a common data structure for allowing simple reuse of output generated by previously executed plugins. Annotations on language resources are used as common interface between individual processing resources. Each processing resource can read and evaluate annotations added to a language resource by a previously executed processing resource. From a developer's point of view this means that a processing pipeline customized to the needs of a specific NLP application can easily be created, modified and executed for a single document or a collection of documents.

The second major advantage of using GATE for developing NLP applications is the availability of easy to use visual resources through which the annotations created by the processing resources can easily be viewed and verified.

NLP with GATE

GATE provides a family of NLP resources grouped together under the name ANNIE (A Nearly-New Information Extraction system) [Gat, Section 6]. The plugins communicate exclusively over annotations on GATE documents and have to be applied in a valid order since some resources are dependent on annotations created by other plugins. ANNIE contains a variety of processing resources from which the tokenizer, sentence-splitter and the part-of-speech-tagger are used in the method developed in this work. Another plugin frequently used in the decomposition rules is GATE's noun phrase chunker [Gat, Section 17.2].

Tokenizer The tokenizer splits a text into the linguistic units it is composed of by identifying simple tokens. It assigns two types of annotations: *Token* and *SpaceToken*.

The *Token* annotation is assigned to each identified unit of text. It contains an attribute called *kind* for differentiating between word, number, punctuation and symbol tokens.

- **Word** – A word is any set of continuous upper or lower case letters. A word may include an apostrophe like in “don’t” but no other punctuations.
- **Number** – Numbers are defined as a sequence of consecutive digits. The *Token* annotation is assigned to the entire sequence rather than to each single digit. So for example the sequence “978” is assigned a single *Token* annotation.
- **Punctuation** – The tokenizer differentiates between three types of punctuations. Punctuations such as “(” and “[” are defined as *start_punctuations*, their counterparts “)” and “]” as *end_punctuation* and punctuations like semicolon and comma as *other punctuations*.
- **Symbol** – The tokenizer differentiates between currency symbols such as “\$” and other symbols like “&”. Any sequence of symbols is assigned a single *Token* annotation.

The white spaces between the identified *Tokens* are marked with a *SpaceToken* annotation.

Sentence-Splitter The sentence-splitter identifies and annotates sentences in a text based on the annotations created by the tokenizer. It assigns a *Sentence* annotation to each identified sentence and a *Split* annotation to each sentence break.

Part-of-Speech Tagger POS-tagging is the process of assigning each word its appropriate category or POS-tag such as “noun” or “verb” based on both the definition and the context of the word. Since basic word categories such as noun are usually not detailed enough for NLP applications, POS-taggers are usually able to differentiate between a large number of subcategories. GATE’s POS-tagger [Hep00] which is a modified version of the Brill tagger assigns to each *Token* annotation one of 54 possible POS-tags which are described in the GATE user guide [Gat, Appendix G]. Since the POS-tagger

processes one sentence at a time and adds the POS-tags to the *Token* annotations, the tokenizer as well as the sentence-splitter have to be executed before the POS-tagger.

Noun Phrase Chunker Text chunking is the process of finding non overlapping text segments based on superficial analysis. The goal of noun phrase chunking (NP-chunking) is to identify sequences of words which together form a noun phrase. GATE’s NP-chunker is an implementation of Ramshaw and Marcus BaseNP chunker [RM95]. The chunker requires texts to be annotated with POS-tags which are evaluated for finding the noun sequences. Each identified noun sequence is assigned a *NounChunk* annotation. For providing a better understanding of NP-chunking the following example shows the noun phrases found in a dependent claim. Each identified noun phrase is enclosed in square brackets.

Example 4.2 *EP1398061-A2*

[A saliva ejector] according to claim [1], wherein [the inlet element] defines [a spherical shape] and is connected to [the disk].

JAPE

With JAPE (Java Annotation Patterns Engine) GATE provides a very powerful method for identifying regular expressions in annotations. A detailed description of JAPE can be found in the GATE user guide [Gat, Section 8]. A developer can write JAPE grammars which evaluate annotations created by previously executed processing resources like the tokenizer or POS-tagger. GATE contains a plugin for transducing and executing these JAPE grammars. A JAPE grammar is composed of one or more phases which are executed sequentially. Each phase consists of a set of pattern/action rules for finding patterns and executing appropriate actions. The idea of JAPE grammars can easily be understood when looking at a concrete example. The simplified example provided in Listings 4.1 and 4.2 shows a JAPE grammar for annotating noun sequences connected with the word “and” such as “patent claims and descriptions”. The described grammar only serves the purpose of providing an overview of the functionality of JAPE and is therefore much simpler than most grammars developed for complex NLP applications.

The grammar consists of two phases. In Listing 4.1 the first phase is shown in which noun sequences are annotated by evaluating POS-tags assigned to *Token* annotations. Line 1 specifies the name of the phase. Line 2 enumerates all annotations which are evaluated by the grammar. In this case

only the *Token* annotations are evaluated while all other types of annotations are ignored. Line 3 provides information on how the rules should be applied. The control value “brill” states that only the longest match should be annotated. Lines 5-10 show the pattern/action rule for annotating the noun sequences. Lines 5-9 are the left-hand side of the rule used to identify the pattern. The actual pattern, which matches a sequence of one or more nouns, is specified in line 7 as regular expression over the attribute *category*, which contains the POS-tag assigned to the *Token* annotation. Line 10 is the right-hand side of the rule where a *NounSequence* annotation is assigned to the pattern identified by the left-hand side of the rule.

```

1 Phase: markNounSeq
2 Input: Token
3 Options: control=brill

5 Rule: markNounSeq
6 (
7   ({Token.category =~ "NN"})+
8 )
9 :nounSequence
10 -->:nounSequence.NounSequence= { rule = "markNounSeq"}
```

Listing 4.1: JAPE grammar: Mark Noun Sequence

In phase two the annotations created in phase one can be used in the left-hand side of pattern/action rules. Listing 4.2 shows how the *NounSequence* annotation can be used to identify noun sequences connected by the word “and”. Since the *NounSequence* annotation is used in the left-hand side of the rule it has to be added as input annotation in line 2 as well.

```

1 Phase: markConjNounPhrase
2 Input: Token NounSequence
3 Options: control=brill

5 Rule: markConjNounPhrase
6 (
7   {NounSequence}
8   {Token.string == "and"}
9   {NounSequence}
10 )
11 :conjNounPhrase
12 -->:conjNounPhrase.ConjNounPhrase= {rule = "
    markConjNounPhrase"}
```

Listing 4.2: JAPE grammar: Mark Conjoined Noun Phrase

4.4 Regular Expressions

Regular expressions are used in this work for text manipulation and text pattern matching as well as for describing JAPE grammars. The regular expressions used for describing JAPE grammars are provided in an informal syntax. The POS-tags listed in Table 4.5 are often used in these regular expressions. In addition to these POS-tags the place holders described in Table 4.6 are used for indicating a set of POS-tags.

Part-of-Speech tag	Meaning
CD	Cardinal number
DT	Article including “a”, “an”, “every”, “no”, “the”, “another”, “any”, “some”, “those”
IN	Preposition or subordinating conjunction
JJ	Adjective, not including comparatives and superlatives
NN	Noun - singular
NNP	Proper Noun - singular
NNPS	Proper Noun - plural
NNS	Noun - plural
RB	Adverb, not including comparatives and superlatives
TO	The word “to”
VB	Verb - base form: subsumes imperatives, infinitives and subjunctives
VBD	Verb - past tense
VBG	Verb - gerund or present participle
VBN	Verb - past participle
WDT	Wh-Determiner such as “which” or “what”

Table 4.5: POS-Tags used in Method Description

The regular expressions used for text matching are provided in the Java regular expression syntax⁸. In Java regular expressions several predefined character classes and boundary matchers can be used, from which those relevant for the description of the method developed in this work, are shown in Table 4.7.

⁸<http://java.sun.com/javase/6/docs/api/java/util/regex/Pattern.html>

Place-holder	Meaning
<i>NOUN</i>	All types of nouns (NN, NNP, NNPS, NNS)
<i>NOUN-PLURAL</i>	All nouns in plural (NNPS, NNS)
<i>VERB</i>	All types of verbs (VB, VBN, VBG, VBD)
<i>WORD</i>	All <i>Token</i> annotations assigned to words
<i>PUNCT</i>	All <i>Token</i> annotations assigned to punctuation symbols
<i>NO-PUNCT</i>	All <i>Token</i> annotations not assigned assigned to punctuation symbols

Table 4.6: Place-Holders used in Method Description

Character class	Meaning
<code>\s</code>	Any white space character
<code>\W</code>	Any non word character, meaning all characters except digits and letters
<code>\A</code>	The beginning of the input
<code>\z</code>	The end of the input

Table 4.7: Regular Expression Character Classes

4.5 Document Parsing

XML Transformation The data sets are represented in a highly structured XML format developed by Matrixware which provides tags for marking up all relevant parts of a patent document. Since the goal of the method developed in this work is the decomposition of patent claims, only the claim section and for informational purposes the title of the invention are extracted. It has to be taken into account that both the claims as well as the title are available in three languages: English, German and French. An XSLT stylesheet is used to extract only the English versions while the German and French translations are discarded together with all other non relevant sections of the patent document.

Example 4.3 shows the structure of an XML document after applying the XSLT stylesheet. An important detail is how the claims are stored in the XML document. Each claim is enclosed in a separate `<claim>` tag making it easy to identify the individual claims of a patent. The `<claim>` tag contains a `<claim-text>` tag which marks up the actual text of the claim. The claim text itself is usually not stored completely unstructured meaning that individual parts of a claims, for example each element of an invention, are contained

in a separate `<claim-text>` element. This leads to a nested structure of `<claim-text>` tags, as shown in the first claim in Example 4.3.

Example 4.3 Transformed XML Document

```
<?xml version="1.0" encoding="UTF-8"?>
<patent-document>

  <invention-title lang="EN" status="new">
    Title of the invention
  </invention-title>

  <claims lang="EN" status="new">
    <claim num="01">
      <claim-text>
        Claim-Text
      <claim-text>
        Claim-Text
      </claim-text>
      Claim-Text
    </claim-text>
    Claim-Text
  </claim>
  <claim num="02">
    <claim-text>
      Claim-Text
    </claim-text>
  </claim>
</claims>

</patent-document>
```

The `<claim>` tag contains an attribute called *num* which usually contains a single number indicating the number of the claim in the document such as shown in Example 4.3. The *num* attribute is, however, not used consistently throughout the patent documents and is very frequently set to “XX”. In some particular cases it also specifies a range in a form such as *num*=”01-05”. This occurs if a patent contains several canceled claims as shown in Example 4.4.

Example 4.4 Transformed XML Document with Canceled Claim

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<patent-document>
```

```
    [...]
```

```
    <claims lang="EN" status="new">
```

```
        <claim num="01-05">
```

```
            <claim-text>
```

```
                canceled
```

```
            </claim-text>
```

```
        </claim>
```

```
        [...]
```

```
    </claims>
```

```
</patent-document>
```

XML Parsing Before a patent document can be processed it has to be parsed from the XML document into main memory. An internal data structure is used which provides an object-orientated representation of the XML document structure. Each patent is represented as a Java object containing a title and a list of claim objects in the same order as in the original XML document. The claim objects contain the claim number and the actual claim text. If a `<claim>` tag contains a valid *num* attribute, the attribute value is parsed and used as claim number. Otherwise the claims are numbered according to their order in the XML document.

For extracting the claim text the structural information present in the XML document is not used meaning that if a patent claim contains nested `<claim-text>` elements, these parts are merged into a single sentence. The reason for this is to keep the developed decomposition rules independent of a given data representation. In addition to that the structural information for claims is not consistent in the document collection and does not contain any information about how the sub-parts are related. If the claim length is less than 100 characters and it contains the word “cancel”, the claim is marked as canceled claim and is ignored by the decomposition algorithm.

4.6 Data Cleaning

Several preprocessing and data cleaning steps have to be executed before the extraction and splitting rules can be applied. The processing resources are implemented as separate Java classes and are executed for the text of each claim extracted from the XML document. In the following sections the purpose and functionality of each preprocessing resource is described in detail.

Image Link Remover In patent claims references to images are enclosed in parentheses. Their representation can include numbers as well as letters and range from simple forms such as “(21)” or “(12b)” to more complex constructs like “(21b;23;25c)”. In the developed method these image links are not processed and pose problems for the extraction rules. The following regular expression is used for finding and removing image links:

```
(\\s*[0-9][0-9a-z,;\\s]*\\))
```

It has to be taken into account that abbreviations, mathematical and chemical formulas may also be enclosed in parentheses. Therefore the first character after the opening parenthesis must be a digit. The digit can be followed by other digits and lower case letters separated by a comma or a semicolon. The reason for allowing only lowercase letters is to make sure that chemical formulas are not affected by the preprocessing resource. If uppercase letters were allowed, chemical formulas like “(2Ca)” would be removed. Mathematical formulas are not affected by the regular expression since it does not match any mathematical operators.

Other information which is contained in parentheses concerns the explicit enumeration of steps of a method or elements of an invention occurring in some claims. This is shown in Example 4.5 where each step of a method is introduced and identified by a lowercase letter enclosed in parentheses. This information is also used later in the claim for referencing an already introduced step. Therefore this information can usually not be removed without a severe impact on the readability of the claim. It is, however, not required or used by the splitting and extraction rules themselves.

Example 4.5 EP1442755-B1

A method of modifying a ceramic-coated implantable article, which method comprises:

- (a) providing an implantable article [...],
 - (b) incubating at least a portion of the bioactive ceramic coating [...],
 - (c) removing the liquid carrier from the bioactive ceramic coating to yield a modified implantable article with a bioactive ceramic coating into which the biological agent is incorporated [...],
- provided that, either (A) the concentration of the calcium ions in the composition used in step (b) is 0.01 to 1.0 mM [...]
-

Enumeration Symbol Remover In some claims, elements of an invention are enumerated in a form such as “a.” or “b.”. This is shown in Example 4.6. Since a period (“.”) occurring in this context is interpreted as a sentence delimiter by GATE’s sentence-splitter these constructs lead to erroneous decomposition of claims. Thus the regular expression

$$(\backslash s+[a-zA-Z]\backslash.\backslash s+)$$

is used for finding and removing all single letters followed by a period. The white spaces before and after the pattern are required in order to ensure that the last letter and the period ending the sentence are preserved.

Example 4.6 EP0171002-B1

A method of forming a dental impression [...] comprising the steps of:

- a. placing a predetermined quantity of the flowable material in a transparent impression tray having walls defining a recess
 - b. impressing the material in the tray against dental anatomy in an oral, cavity of which a mold is desired
-

Claim Number Remover In many documents the actual claim text is preceded by its claim number. This can be seen in the XML patent document shown in Example 4.7. Since this information is already implicitly given via the order of the claims in the patent document it is removed via the regular expression

$$(\backslash A\backslash s*[0-9]+\backslash s*[\backslash .,:\backslash \)])?$$

which finds sequences of digits followed by a period, a colon, a comma or a closing parenthesis from the beginning of an input text.

Example 4.7 EP0411105-B1: Claim Numbering

```
<claims lang="EN" status="new">
  <claim num="xx">
    <claim-text>
      1. The use of a salt of lanthanum for [...]
    </claim-text>
  </claim>
  <claim num="xx">
    <claim-text>
      2. The use of lanthanum chloride for [...]
    </claim-text>
  </claim>
</claims>
```

Characterized Spelling Normalizer Due to spelling differences in American English and British English the keyword “characterized” is sometimes spelled with a “z” and sometimes with an “s”. For this particular keyword the spelling needs to be normalized since it needs to be identified correctly in a number of processing steps. The preprocessing resource replaces all occurrences of the word “characterised” with “characterized”.

Additionally the preprocessing resource corrects an error occurring in a small number of XML documents. In some patent claims the white spaces between the keywords “characteri[sz]ed in that” or “characteri[sz]ed by” are missing, thus they are written as “characteri[sz]edinthat” or “characteri[sz]edby”. In these cases the Characterized Spelling Normalizer inserts the appropriate white spaces and normalizes the spelling in a single step.

Said Replacer This processing resource replaces all occurrences of the word “said” with the definite article “the”. This is a simple way of improving the performance of natural language parsers even before decomposing the claims. Natural language parsers trained on general language texts interpret the word “said” as verb. In claims, however, it is always used for referring to an already introduced concept.

4.7 Claim Type Identification

4.7.1 Dependent or Independent Claim Identification

A simple heuristic is used to determine whether a claim is a dependent or an independent claim. The drafting guidelines for dependent claims (described in Section 2.6.3) suggest that a dependent claim should consist of two parts. The first part contains a reference to the claim or claims which are refined written in a form such as “The dental handpiece of **claim 1**” or “The orthodontic bracket of any one of **claims 1 to 7**”. All claims containing either the word “[Cc]laim” or “[Cc]laims” (with the bracket indicating that letter can be written in lower as well as upper case) are classified as dependent claims. All other claims are treated as independent claims. The advantage of this approach is its simplicity and transparency. The major disadvantage is that it does not take into account that in rare cases also independent claims may contain a reference to other claims. The reasons why an independent claim may refer to other claims are examined in Section 2.6.1. In some cases independent claims are therefore erroneously classified as dependent claims.

4.7.2 Claim Category Identification

Within independent claims it is differentiated between three categories: physical entity claims, method claims, use claims. A simple heuristic based on keyword matching is used for this purpose. Since the developed method is based on linguistic patterns found in claims and does not deal with any legal aspects, the defined categories may differ from the categories commonly used in the IP domain.

Method Claim Identification The examination of the Analyzed Set has shown that claims containing the keyword “method” or “process” within the first 100 characters can be classified as method claims. Two such claims are shown in Examples 4.8 and 4.9.

Example 4.8 EP1442755-B1

A method of modifying a ceramic-coated implantable article, which method comprises: [...]

Example 4.9 EP1442755-B1

A process for producing a magnet structure comprising steps of: [...]

Use Claim Identification All claims which start with the phrase “The use” are classified as use claims. Thus simple string matching can be used to classify these claims accordingly. Only very few use claims are available in the Analyzed Set. One such claim is shown in Example 4.10

Example 4.10 EP0187757-B1

The use of potassium bicarbonate for the manufacture of a composition for [...]

Physical Entity Claim Identification No such simple heuristics are available for identifying physical entity claims. This is mainly due to the fact that physical entity claims start with the claimed invention rather than with claim-specific keywords. Physical entity claims are therefore considered the default case meaning that claims that can neither be classified as use claims nor as method claims are classified as physical entity claims.

Frequency of Claim Categories Tables 4.8 and 4.9 show the frequency of each claim category in the two data sets. The figures show that the number of physical entity claims is almost three times higher than those of method claims in the Analyzed Set and almost nine times higher in the Evaluation Set. They also show that almost no use claims are present in the data sets.

Claim type	Number of claims
Physical Entity Claims	114
Method Claims	41
Use Claims	4

Table 4.8: Analyzed Set: Claim Types

Claim type	Number of claims
Physical Entity Claims	10,310
Method Claims	3,315
Use Claims	3

Table 4.9: Evaluation Set: Claim Types

4.8 Claim Decomposition

The process of decomposing a claim into smaller parts consists of three main phases. The extraction process can be applied to the entire original claim text or can be executed for already extracted parts for decomposing them further into even smaller units. The main phases are pattern identification, pattern extraction, post processing and merging the extracted parts into a tree structure.

4.8.1 Pattern Identification

Some patterns can be identified through simple lexical matching of keywords. If this is possible, patterns are identified using Java regular expressions. Most patterns, however, are more complex and thus require deeper linguistic analysis of the claim. Each claim is tokenized and a sentence-splitter is applied. Depending on the requirements of the extraction rules POS-tagging and NP-chunking is done. A small JAPE grammar is used to extend the annotations created by GATE's sentence-splitter and tokenizer. It adds an attribute named *delimiter* with the values "start" and "end" to the first and last *Token* annotation in each sentence. The created linguistic information is used in claim-specific JAPE grammars for identifying and marking extractable patterns. These grammars are described in detail in the Sections 4.9 and 4.10.

4.8.2 Pattern Extraction

Based on the annotations created by the JAPE grammars the claims can be decomposed. For this purpose the actual text content of each annotated pattern is extracted from GATE's internal flat document representation into a GATE independent hierarchical tree data structure. For each extracted part a number of post processing steps are executed.

Post Processing

The post processing resources remove unnecessary characters such as white spaces and punctuation symbols and unnecessary words from the extracted parts.

White Space Normalizer The White Space Normalizer applies two regular expressions to a given string for removing unnecessary white spaces. The first regular expression replaces several subsequent white spaces with a single white space. The second regular expression removes blanks occurring

before a comma. Superfluous white spaces may already be contained in the original document or may be introduced during the splitting and extraction procedure.

Word and Punctuation Remover Many extracted parts contain unnecessary punctuations and words at the beginning and the end of a sentence. This occurs mainly for composition-parts for which the semicolon or comma used for separating one part from another is extracted together with the text. In some cases unnecessary punctuation symbols also occur at the beginning of an extracted part. Nested sentences for example are extracted with the comma preceding the keyword introducing the pattern. Many extracted parts start or end with the word “and”. This happens mainly for elements of an invention extracted from Composition-Patterns. Since these elements are “and” connected implicitly this word can be removed. The two regular expressions

$$(\backslash A((\text{and})?[\backslash W\&\&[\wedge\backslash ()] ?)*)$$

and

$$(((\backslash s+\text{and})?[\backslash W\&\&[\wedge\backslash ()]] ?)*)\backslash z)$$

are used to remove the word “and” and all non-word characters, except opening and closing parentheses, from the beginning and the end of each part. Parentheses can not be removed since they are sometimes found at the beginning of an extracted part for identifying steps of a method in a from such as “(a)” or “(b)”.

Internal Data Structure

The decomposed claims are stored in a tree structure. Each node in the tree contains an extracted part of the claim. The edges represent the type of relation to the parent. Each node contains the text of the extracted part and, in order to be able to traverse the tree, a reference to its parent relation and a list of child relations. Each relation contains an enumerated type indicating the type of the relation and an optional string containing a label for the relation. It also contains a reference to its parent and a list of references to its child nodes.

4.9 Independent Claim Decomposition

The main focus of this work is the decomposition of independent claims since they are much longer and more complex than dependent claims. Due to large structural differences of claims from different categories only a very limited number of rules which are applicable to all claim types is available. The major part of the developed rules is specific to one of the claim categories.

4.9.1 General Patterns

Before a claim is decomposed using the claim category-specific rules the following two patterns are extracted.

Claim-Subject

A claim-subject is extracted and used as the root node of the tree structure. The claim-subject is that part of the claim to which all other claim parts are directly or indirectly related to. For method and use claims the identification of the subject is rather trivial. In method claims all other extracted parts can be attached to the initial keyphrase “A method” or “A process”. For use claims they can be attached to the phrase “The use”. While the claim-subject for these two categories could be extracted using a simple string matching approach, this is usually not the case for physical entity claims. In physical entity claims the root of the sentence is the invention itself. This is illustrated in Example 4.11. Therefore each claim sentence is analyzed with GATE and the first noun phrase is extracted as claim-subject.

Example 4.11 EP1444966-A1

Claim-Subject

A dental head unit capable of measuring a root canal length of a patient

Characterized-Pattern

If a claim is drafted in the two-part form suggested by the EPO, the keyphrases “characterized in that” and “characterized by” can be used to split the claim into two parts, the preamble and the claim body. This pattern can be exploited without linguistic analysis. Regular expressions are used to split the claim text where either of the keyphrases mentioned above occurs. The characterized-part (claim body) is attached to the root of the tree structure with a CHARACTERIZED relation. For physical entity claims the

characterized-part is further analyzed with the rules described in Subsection “Characterized-Part Decomposition” of Section 4.9.2. The preamble itself is not attached to the tree structure. It is decomposed using the category-specific rules described in the following sections. If a claim does not contain a Characterized-Pattern, the entire claim text is decomposed using these claim category-specific rules.

4.9.2 Physical Entity Claims

The focus in this method was set on the analysis of physical entity claims. Due to the – compared to the other claim categories – large number of physical entity claims in the Analyzed Set it was possible to identify a larger number of patterns. Additionally the structure of physical entity claims lends itself very well to rule-based decomposition. Therefore these claims can be split into very small consistent units.

Composition-Pattern

The pattern which occurs most frequently in physical entity claims is the Composition-Pattern since an invention is usually described by enumerating all elements it is composed of. Thus the complexity of claims can be significantly reduced by correctly extracting these elements. The Composition-Pattern is introduced by one of the keywords “comprising”, “comprises” or “including” and is composed of several composition-parts. Each of these composition-parts describes an element of the invention and therefore starts with a concept mentioned in the claim for the first time. The parts can be identified by looking for noun phrases in plural or noun phrases in singular preceded by the indefinite article “a” or “an”. Example 4.12 should provide a better understanding of the Composition-Pattern found in physical entity claims.

Example 4.12 EP0063891-B2

Claim-Subject	
An ejector holder [...]	, the holder comprising
Composition-Start	
Composition-Part	Composition-Part
an elongate barrel[...],	a plunger [...], [...]
Composition-Part	
and a lever [...].	

The JAPE grammar used for extracting Composition-Patterns consists of several phases.

Composition Start In the first phase the JAPE grammar marks the start of a Composition-Pattern. It assigns a *Comp-Start* annotation to patterns matching the regular expression shown in Listing 4.3. The place holder *COMP-WORD* stands for one of the words “comprising”, “including” or “comprises”

```

1  (,)?
2  ( (said | the | each | which) (NO-PUNCT) [0,5] )?
3  (COMP-WORD)
4  (PUNCT)?

```

Listing 4.3: Regular Expression: Comp-Start

Line 1 includes a comma into the *Comp-Start* annotation if it precedes the actual pattern. Line 2 is needed because the claim-subject may be repeated before the *COMP-WORD* in the claim. As the phrase repeating the subject may not be completely identical to the subject itself, it is not possible to check for the exact words used in the claim-subject. Therefore a sequence of zero to five words which are not separated by punctuation characters (*NO-PUNCT*) are allowed before the keyword introducing the pattern. Line 3 then matches the actual keyword and line 4 allows a punctuation symbol at the end of the pattern.

The JAPE grammar is configured to annotate only the longest match and only the first *Comp-Start* annotation in each sentence is retained.

Composition-Parts The second phase identifies and annotates the composition-parts the Composition-Pattern is composed of. The annotation procedure consists of several steps.

Possible-Composition-Part-Start Annotation In the first step all grammatical constructs possibly starting a composition-part are annotated with a *Possible-Comp-Part-Start* annotation. The regular expression shown in Listing 4.4 describes the JAPE grammar used for this purpose. The place holder *CONJ* stands for the conjunctions “and” and “or”. The place holder *ENUM-MARKER* matches a single *Token* annotation enclosed in parentheses such as “(a)” or “(iv)” as well as *Token* annotations followed by a closing parenthesis such as “a”.

```

1  (CONJ)?
2  (ENUM-MARKER)?
3  (
4    (a | an) |
5    (NOUN-PLURAL) |
6    ( (WORD ^ (DT,each)) [0,5] (means) ) |
7    (at least)
8  )

```

Listing 4.4: Regular Expression: Possible-Comp-Part-Start

Lines 3 to 8 form the core part of the grammar describing valid patterns possibly introducing a composition-part. Line 4 identifies noun phrases in singular by matching the indefinite articles “a” and “an”. Since an indefinite article is always followed by a noun phrase in singular the rule does not look for POS-tags to avoid problems with adjectives preceding the noun phrase. Line 5 annotates noun phrases in plural while line 6 implements a special rule for the word “means”. This special rule is necessary to compensate for incorrect POS-tagging of the word “means” which is tagged as a verb and not as a noun by GATE’s POS-tagger. Always considering the word “means” a noun, as done in the rule above, does not lead to incorrect results. This is due to the special characteristics of patent claims where the use of “means” as a verb is very unlikely. Since the word “means” is often preceded by an adjective the rule allows the keyword to be preceded by zero to five words excluding determiners and the word “each”. Line 7 matches the keyphrase “at least” which is also commonly used for introducing a composition-part.

Filtering In a second step the *Possible-Comp-Part-Start* annotations which do not start a composition-part are filtered out. The others are marked with a *Comp-Part-Start* annotation. A *Possible-Comp-Part-Start* annotation is considered a correct start of a composition-part if it is preceded by a *Comp-Start* annotation, a comma or a semicolon. An attribute called *type* with the values “comma”, “semicolon” or “comp-start” is used in the *Comp-Part-Start* annotation to retain the information about the preceding character or annotation.

Composition-Part Annotation In the final step, after all *Comp-Part-Start* annotations are found, a JAPE grammar is used to annotate the complete composition-parts which can then be extracted. A composition-part starts after a *Comp-Part-Start* annotation and ends when either another *Comp-Part-Start* annotation is found or the sentence ends. The JAPE gram-

mar annotates composition-parts separated by semicolons with a different annotation (*Comp-Part-Semicolon*) than those separated by commas (*Comp-Part-Comma*). This is necessary because one Composition-Pattern can be nested inside another. If this is the case, *Comp-Part-Comma* annotations can occur inside of *Comp-Part-Semicolon* annotations. Since nested compositions are not extracted the algorithm first scans for *Comp-Part-Semicolon* annotations and only if none are found *Comp-Part-Comma* annotations are extracted. The extracted parts are attached to the claim-subject node with a COMPOSITION relation.

Nested-Sentence-Pattern

Since each claim has to be written in one sentence, certain grammatical structures are used for chaining separate sentences in order to create one single sentence. A very common structure used for this purpose is the Nested-Sentence-Pattern where a concept, which has already been introduced, is refined. Example 4.13 should provide a better understanding of the Nested-Sentence-Pattern.

Example 4.13 EP0028529-B2

Claim-Subject

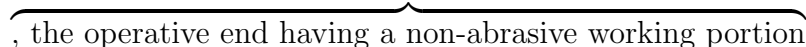
A scaler tip  having an operative end

Description

Nested-Sentence-Part

, the operative end terminating in a curved free end

Nested-Sentence-Part

, the operative end having a non-abrasive working portion

There are several, very similar, keyphrases which introduce a nested sentence. The phrases “, the CONCEPT” or “; the CONCEPT” where CONCEPT represents an already introduced concept are used frequently. In the original claims the word “said” is often used instead of the article “the”. However, since all occurrences of the term “said” are replaced by the word “the” during the preprocessing steps only the keyword “the” has to be taken into account. The two patterns above occur in slightly modified versions containing the words “and” and “wherein”. A JAPE grammar is used to mark the beginning of each Nested-Sentence-Pattern according to the following regular expression:

`(,|;) (and)? (wherein)? (the)`


A nested sentence ends when either another Nested-Sentence-Pattern is found or the sentence ends. The rule for annotating nested sentences partly overlaps with the rule for annotating the start of a Composition-Pattern. Therefore the process of annotating and extracting nested sentences is started after the extraction of the Composition-Pattern. The extracted sentences are attached to the claim-subject node with a NESTED-SENTENCE relation.

Description-Pattern

All words between the claim-subject and the first pattern which is found in the claim (Nested-Sentence or Composition-Pattern) are extracted as description-part. The description usually indicates the purpose of the invention such as shown in Example 4.14. In some cases it may, however, also describe elements an invention contains.

Example 4.14 EP0415508-A2

Claim-Subject

An apparatus  comprising [...]

A simple JAPE grammar is used to annotate all words after the claim-subject with a *Description* annotation until either a Nested-Sentence-Pattern or a Composition-Pattern is found or the claim sentence ends. The part annotated with a *Description* annotation is extracted and appended to the claim-subject node in the data structure with a DESCRIPTION relation.

Characterized-Part Decomposition

If the claim is drafted in the two-part form suggested by the EPO, the characterized-part extracted with the Characterized-Pattern rule can be decomposed further into smaller parts. The annotation and extraction process first looks for extractable enumerations of elements. For this purpose the already described Composition-Pattern rules are used in a slightly modified version. The only difference lies in the JAPE grammar used for annotating the start of a Composition-Pattern. The rule is much simpler as shown in Listing 4.5.

```
1  (COMP-WORD)
2  (PUNCT)?
```

Listing 4.5: Regular Expression: Comp-Start

Line 1 matches one of the keywords “comprising”, “including” and “comprises” while line 2 allows an optional punctuation symbol after the keyword. The rules used for annotating the composition-parts remain the same. In the Java code, however, the pattern is only extracted when more than one composition-part is found. Otherwise the Composition-Pattern is not extracted. The extracted parts are attached to the node containing the characterized-part with a COMPOSITION relation.

Parts of an invention specified in the characterized-part are not necessarily enumerated using a Composition-Pattern. In some cases the parts are simply separated from each other by semicolons. Therefore if no Composition-Pattern is found, the characterized-part is simply split by semicolons. If this results in more than one part, each of these parts is added to the node containing the characterized-part with a CHARACTERIZED-COMPOSITION relation.

Composition-Part Decomposition

Extracted composition-parts can be decomposed further by splitting them into a part containing the element of the invention and a second part containing a description of the element. This is illustrated in Example 4.15.

Example 4.15 EP1484028-A2

Element-Part	Description-Part
[...] a chuck assembly	secured to the rotor shaft
Element-Part	Description-Part
[...] a positioning template	for guiding the positioning and bonding [...]

A JAPE grammar is used to identify the end of the element-part by looking for specific linguistic patterns. Some of these patterns make use of a *NounSequence* annotation created by a previously executed JAPE grammar using the following regular expression:

$((\text{NOUN} \mid \text{JJ})^* \text{NOUN})$

The following patterns, used for ending the element-part, are marked with an *Element-Part-End* annotation:

- $(\text{RB})? (\text{VBG}) !(\text{NounSequence})$

The pattern matches a verb in gerund form possibly preceded by an adverb. This is a very frequently occurring pattern as shown for example

in the phrase “a neck section **extending proximally** from the head section [...]”. The reason for not allowing the verb in gerund form to be followed by a *NounSequence* annotation is to make sure that phrases like “a teeth engaging element” are not split into the two parts “a teeth” and “engaging element”. The optional adverb before the verb is used for correctly splitting phrases like “a first bearing assembly **radially supporting** the motor assembly”.

- (having | comprising | including)
This is essentially the same rule as above. But as opposed to other verbs in gerund form the keywords “comprising”, “including” and “having” are never used in compound nouns. Therefore they end the element-part also when they are succeeded by a *NounSequence* annotation.
- (RB)? (VBD | VBN) !(*NounSequence*)
This pattern matches a verb in past tense, possibly preceded by an adverb. It is used for descriptions of parts written in passive form such as “a brush part **detachably attached** to one end of the drive shaft”. The !(*NounSequence*) is needed because otherwise certain phrases would be erroneously split into two parts, like the phrase “a wedged body” into “a wedged” and “body”.
- (JJ) (TO | IN)
The description-part is often separated from the element-part by phrases such as “configured to” or “connectable at”. This is illustrated in the sentence “an ejector passage tube **connectable at** a first end to a suction source”.
- (for | each)
Often the words “for” or “each” are used to separate the two parts such as in the sentence “driver means **for** applying reciprocal force to the drive shaft.”
- (WDT)
This pattern matches so called wh-determiners like “which” or “that”, like in the phrase “a keeper body **which** is shaped like a plate”.
- (TO) (VB)
This pattern describes phrases in which the two parts are split by the preposition “to” followed by a verb in base form, such as in the sentence “a second end connection **to connect** to an aspiration unit”.

- (*PUNCT* except “(” and “)”)

This pattern matches all punctuation symbols except opening and closing parentheses. It is used for splitting phrases such as “spatial orientation means , with which it is possible to [...]”. Opening and closing parentheses are excluded to take into account that a composition-part can be started with a phrase such as “(a)”.

In the Java code all *Token* annotations starting from the *Element-Part-End* annotation are extracted as composition-part-description. The element-part remains in the already existing composition-part node. The extracted description is added to it with a COMP-PART-DESCRIPTION relation. The description-part itself can be decomposed into even smaller units by extracting nested sentences. This is done with the already described Nested-Sentence-Pattern rule.

4.9.3 Method Claims

Method claims are the second largest group of claims in the Analyzed Set. The patterns found in methods claims are similar to those found in physical entity claims. However, the keyphrases and grammatical structures used in the patterns are different. Therefore a separate rule set is needed for method claims.

Composition-Pattern

The Composition-Pattern of methods is usually a sequence of steps which are performed in order to achieve a certain goal. The keywords for introducing the Composition-Pattern are: “comprising”, “comprises” and “including”. These keywords are then followed by an enumeration of one or more steps. Each step is normally introduced by a verb in gerund form. Example 4.16 should provide a better understanding of Composition-Patterns found in method claims.

Example 4.16 Composition-Pattern in Method Claims

Claim-Subject

$\overbrace{\text{A method [...] comprising the steps of:}}^{\text{Composition-Start}}$
 $\overbrace{\text{preparing an [...]}}^{\text{Composition-Part}}; \underbrace{\text{placing a [...]}}_{\text{Composition-Part}}; \overbrace{\text{removing said [...]}}^{\text{Composition-Part}}.$

The JAPE grammar used for extracting these Composition-Patterns is similar to the grammar used for physical entity claims. Only the rules for creating the *Comp-Start* and *Possible-Comp-Part-Start* annotations are different in terms of keyphrases used for identifying the patterns.

Composition Start Several very similar grammatical patterns are used for starting a Composition-Pattern in method claims. The regular expression shown in Listing 4.6 is implemented as a JAPE grammar to annotate the start of a Composition-Pattern with a *Comp-Start* annotation. The place holder *COMP-WORD* stands for one of the words “comprising”, “including” and “comprises”.

```

1  (,)?
2  ((the | which) (method | process) )?
3  (COMP-WORD)
4  ((the | a)? (following)? (step | steps | acts) (of)??)
5  (PUNCT)?

```

Listing 4.6: Regular Expression: Comp-Start

Line 1 includes a possible comma into the pattern. Line 2 takes a common case into account where the claim-subject is repeated before the keyword introducing the composition. Line 3 matches the actual keyword identifying the pattern. Lines 4 and 5 are necessary since it is common that the *COMP-WORD* is followed by phrases like “the steps of:” before the enumeration of the actual steps starts. As for the Composition-Pattern in physical entity claims only the longest match is annotated and only the first *Comp-Start* annotation in each sentence is retained.

Composition-Parts After the *Comp-Start* is annotated each step of the method is extracted as a composition-part. For this purpose a JAPE grammar is used to annotate all verbs in gerund form except *COMP-WORDS* with a *Possible-Comp-Part-Start* annotation. The JAPE grammar allows the verb in gerund form to be preceded by an adjective and/or an enumeration marker such as “(a)”. This is necessary to correctly annotate phrases such as: “(b) **rotatably indexing** the instrument blank [...]”.

After the parts are annotated the same filtering and extraction procedure as used for composition-parts of physical entity claims is applied. Each extracted part is attached to the claim-subject node with a COMPOSITION relation.

Description-Pattern

The claim-subject is normally followed by a description of the method. In the Analyzed Set the words “for”, “of” and “to” were identified as keywords introducing a Description-Pattern. The extracted description-part starts with the first phrase after the keyword and ends when either a Composition-Pattern is found or the claim sentence ends. Example 4.17 should help to better understand the applied rules.

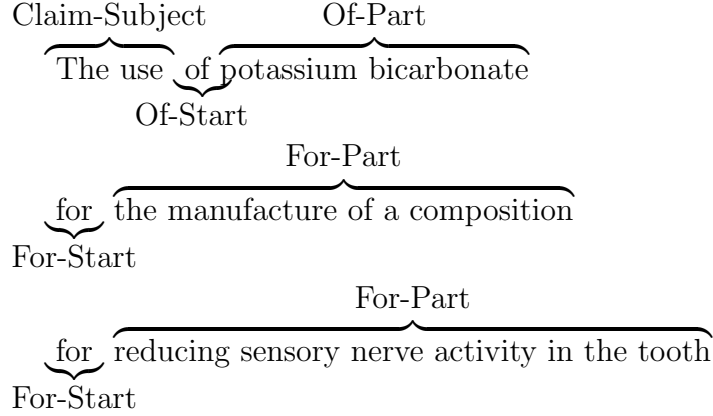
Example 4.17 EP0154137-B1

Claim-Subject	Description
A method	for preparing a dental restoration, (Composition-Pattern)
Description-Start	Description-End

A JAPE grammar is used that performs the annotation in two phases. In the first phase the words “for”, “of” and “to” are annotated with a *Descr-Start* annotation if they are preceded by the word “method” or “process”. In the second phase all words following the *Descr-Start* annotation are marked with a *Description* annotation until either a Composition-Pattern is found or the sentence ends. The annotated part is then extracted and appended to the claim-subject node with a DESCRIPTION relation.

4.9.4 Use Claims

Finding generic rules for decomposing use claims was difficult due to the small number of independent use claims in the Analyzed Set. Nevertheless the structure of the use claims available is very clear and simple compared to the structure of claims from the other two categories. Use claims start with the keywords “The use” followed by the keyword “of” which introduces the description of the material or apparatus used. Additionally the use claims in the Analyzed Set contain one or more parts describing the purpose for which the material or apparatus is used. These parts are introduced with the keyword “for”. Example 4.18 should help to better understand the structure of use claims.

Example 4.18 EP0187757-B1

A JAPE grammar is used to annotate the parts in two phases. In the first phase the word “of” is annotated with an *Of-Start* annotation if it follows directly after the claim-subject. Each occurrence of the word “for” is annotated with a *For-Start* annotation. In the second phase all words following the *Of-Start* annotation are marked as of-part until a *For-Start* annotation is found. All words following a *For-Start* annotation are marked as for-part until either another *For-Start* annotation is found or the sentence ends. The annotated parts are then extracted and added to the tree structure. The extracted of-parts are attached to the claim-subject node with a USE-OF relation and the extracted for-parts with a USE-FOR relation.

4.10 Dependent Claim Analysis and Decomposition

As analyzed in Section 2.6.3 dependent claims consist of two parts. The first part provides a reference to the claim or claims it refines while the second part describes the refinement itself. Therefore the analysis of dependent claims consists of two tasks. In the first analysis step the reference-part is analyzed to extract the references to refined claims. These references can then be used to assign each dependent claim to all the claims it refines. In the second phase the claim is split into a reference and a refinement-part. For dependent physical entity claims the refinement-part is decomposed with rules similar to the rules used for decomposing independent claims.

4.10.1 Reference Analysis

References are provided in various forms like as a single number, an enumeration of numbers, a range of numbers and sometimes as written text. For each of these cases several rather similar patterns have to be taken into account. A JAPE grammar is used to annotate the following patterns with a *Claim-Reference* annotation.

Range of Numbers If a dependent claim refines several previous claims, the references are often specified as a range of numbers. Two numbers connected by either the keyword “to” or a hyphen (“-”) are used to specify the first and the last claim which the dependent claim refines. The patterns are illustrated in Examples 4.19 and 4.20. A JAPE grammar is used to mark these patterns with a *Range* annotation. In a second phase each *Range* annotation preceded by the keywords “[Cc]laim” or “[Cc]laims” is marked with a *Claim-Reference* annotation.

Example 4.19 EP1442755-B1

An article as claimed in any of claims **12 to 14**, wherein [...]

Example 4.20 EP1488758-B1

The dispensing cartridge of any of claims **1-12**, wherein [..]

Single Number A reference to a single preceding claim, as shown in Example 4.21, is the simplest and most frequently occurring pattern. The JAPE grammar matches the word “[Cc]laim” followed by a number. The number is marked with a *Claim-Reference* annotation if it is not already annotated with a *Range* annotation.

Example 4.21 EP1384449-B1

The locator of **claim 1** wherein [...]

List of Numbers References to more than one claim can also be specified in a comma separated list. This pattern is illustrated in Example 4.22. The JAPE grammar assigns a *Claim-Reference* annotation to a sequences of numbers separated by a punctuation symbol or the keywords “and” or “or”. The pattern is only annotated if none of the numbers has already a *Range* annotation assigned to it.

Example 4.22 EP1520597-B1

The assembly of claim **12, 13, or 14**, wherein [...]

Written Specification A dependent claim sometimes refines all preceding claims. This is usually indicated with phrases such as “according to any preceding claim” or “as claimed in any of the preceding claims”. Examples 4.23 and 4.24 illustrate this pattern by showing two common cases. Such references are annotated with a *Claim-Reference* annotation if they match one of the following patterns:

(any) (WORD)[0,4] (preceding | previous) ([Cc]laim | [Cc]laims)

or

(one) (WORD)[0,4] (preceding | previous) ([Cc]laims)

The pattern is only annotated if it is not followed directly by a number. This is necessary, since otherwise claims containing phrases such as “any one of preceding claims 1-4” would be considered as dependent to all previous claims.

Example 4.23 EP0453493-B1

A syringe according to **any of the preceding claims**, wherein [...]

Example 4.24 EP1609433-A1

The device according to **one or more of the preceding claims**, characterized in that [...]

Combined Specification In some cases the reference to previous claims is provided by specifying a range of claim numbers and additionally an enumeration of numbers. This is illustrated in Example 4.25. In this case the JAPE grammar annotates both references with a *Claim-Reference* annotation.

Example 4.25 EP1354566-A2

The method of any one of claims **1 to 14, 32 or 33** wherein [...]

Reference Extraction

In the Java code the JAPE annotations are extracted and evaluated. Each claim object in the internal data structure is assigned a list of dependent claims based on the extracted claim reference numbers.

4.10.2 Claim Splitting

Reference-Part and Refinement-Part Annotation

In this phase the claim is split into two parts, the reference-part and the refinement-part. The JAPE grammar used for this purpose works in two phases. In the first phase the end of a reference-part is marked with a *Ref-Part-End* annotation. In the second phase all *Token* annotations from the beginning of a sentence to the *Ref-Part-End* annotation are assigned a *Reference* annotation and all words following the *Ref-Part-End* annotation are annotated with a *Refinement* annotation.

Ref-Part-End Annotation Several patterns are used for assigning the *Ref-Part-End* annotation. Since these patterns overlap partly the rules for matching the patterns are fired according to a given priority. In the following description the rules are ordered from the highest to the lowest priority. For each pattern a few examples are provided.

- (*PUNCT*) (wherein | (characterized (in that | by))
This is the most commonly used pattern where the reference-part ends with one of the phrases “, wherein”, “, characterized in that” or “characterized by”. Examples 4.26, 4.27 and 4.28 illustrate these three cases.

Example 4.26 EP0419626-B1

Hinge member as claimed in claim 1, **wherein** the head means is circular [...].

Example 4.27 EP1348387-B1

The device according to claim 7 , **characterized in that** material is surgical steel or titanium.

Example 4.28 EP0455727-B1

Dental anchor of claim 4 , **characterized by** a reduced diameter portion interconnected between the retention portion and the manipulating portion.

- (*PUNCT*) !(or | and | CD)
A punctuation symbol is also used very often for separating the element and the description-part. It is, however, also used in the reference-part itself for enumerating a number of referenced claims, thus the

punctuation symbol only ends an element-part if it is not followed by the keywords “or”, “and” or a number. The pattern is illustrated in Example 4.29.

Example 4.29 EP0474776-B1

An apparatus according to claim 3, 4 or 5 , the reference object comprising an essentially plane plate

- (further)? (comprising)

In some cases the dependent claim does not refine an already introduced concept but refines the whole invention by specifying an additional element which the invention comprises. This is shown in Example 4.30.

Example 4.30 EP1570804-A1

The dental implant of claim 2, **further comprising** one or more of the following means: snap coupling, press fitting [...]

- (wherein | (characterized (in that | by))

This pattern is the same as the first one except that the keywords are not preceded by a punctuation symbol. One such case is shown in Example 4.31.

Example 4.31 EP1384449-B1

The locator of claim 1 **wherein** the stimulus voltage has a single frequency.

- (in) (which)

Sometimes the keywords “in which” are used to separate the reference and the refinement-part such as shown in Example 4.32.

Example 4.32 EP0171002-B1

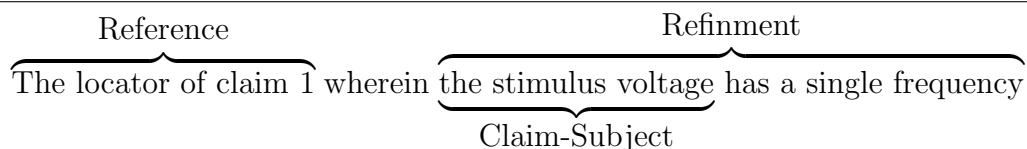
EP0171002-B1 The impression tray according to claim 1 **in which** the light-reflecting means comprises a thin layer of reflective metal.

Dependent-Claim-Subject Extraction

As for independent claims, a claim-subject is extracted as the root node of the tree data structure. For this purpose the first noun chunk in the refinement-part is extracted if it is an already introduced concept. This means that it

either starts with the word “the” or “each”. Example 4.33 should provide a better understanding of the claim-subject extraction rule.

Example 4.33 EP0171002-B1



In two particular cases the subject consists of two connected noun chunks. In several claims an element of an invention is described as “plurality of CONCEPT” or “bundle of CONCEPT” such as in the phrases “the plurality of individual light sources” or “the bundle of fiber optic strands”. The used JAPE grammar takes care that in these cases the complete phrase is extracted as claim-subject and not only the first noun chunk.

If no valid claim-subject can be found, the label of the root element of the tree structure is left empty. The refinement-part is added to the claim-subject node with a REFINEMENT relation, the reference-part with a REFERENCE relation.

4.10.3 Refinement-Part Decomposition

The refinement-parts extracted from dependent physical entity claims are decomposed further by extracting Composition as well as Nested-Sentence-Patterns. The rules for extracting Nested-Sentence-Patterns are the same ones which are used in the decomposition of independent physical entity claims. The Composition-Patterns are extracted with the same grammar used for decomposing characterized-parts from physical entity claims. These rules are described in Section 4.9.2.

4.11 Merging of Dependent and Independent Claims

After the claims have been analyzed and decomposed, an algorithm is applied for merging each independent physical entity claim with its direct and indirect dependent claims. For this purpose the refinement-parts extracted from dependent claims are attached directly to the node in the tree data structure where the refined element was introduced. The idea is illustrated in Example 4.34 which shows the decomposition of an independent and a

dependent claim and in Example 4.35 which shows how the dependent claim can be merged into the tree data structure of the independent claim. The refinement-part “the base member consists essentially of [...]” from the dependent claims is directly attached to the composition-part “a base member” which introduces the refined element in the independent claim.

4.11.1 Merge Process

For attaching refinements from dependent claims to the correct node in the tree structure of the independent claim, the noun phrase introducing the refined element has to be found. For this purpose it can be exploited that a new element is usually introduced with a phrase such as “a CONCEPT” and later referred to as “the CONCEPT”.

Concept Identification

A JAPE grammar is used which annotates, in each claim part extracted from dependent and independent claims, the two types of noun phrases described above. Noun phrases introducing a new element of an invention are marked with a *New-Concept* annotation and noun phrases referencing an already mentioned concept are annotated by the JAPE grammar with a *Ref-Concept* annotation.

Ref-Concept Annotation The grammar used for marking ref-concepts is similar to the grammar used for marking the claim-subject of a dependent claim. It marks all noun chunks starting with either the word “the” or the word “each” with a *Ref-Concept* annotation. In two particular cases a *Ref-Concept* annotation is assigned to two connected noun chunks. This occurs for noun phrases in the form “the plurality of CONCEPT” or “the bundle of CONCEPT”, such as for example in the phrases “the plurality of individual light sources” or “the bundle of fiber optic strands”.

New-Concept Annotation All noun chunks which are not marked as ref-concepts are annotated with a *New-Concept* annotation. The only exception are noun chunks which contain the word “claim”. These phrases are excluded since they are used for referencing previous claims and not for introducing new concepts. Also new-concepts may consist of two noun chunks. The same patterns described for ref-concepts can be applied to new-concepts but instead of the definite article “the” the indefinite article “a” is used. This can, for example, be seen in the phrase “a plurality of dental floss holders”.

Example 4.34 Claims Before Merging

Independent claim:

An oral appliance for placing in a mouth of a user, the appliance comprising: a base member having a generally U-shaped form corresponding to the outline of a jaw of a user, [...]

Subject: An oral appliance

Relation: DESCRIPTION

->for placing in a mouth of a user

Relation: COMPOSITION

->a base member

Relation: COMP_PART_DESCRIPTION

->having a generally U-shaped form corresponding
to the outline of a jaw of a user [...]

Dependent claim:

An oral appliance according to any one of claims 1 to 3, wherein the base member consists essentially of a rigid plastics material which is polyethylene.

Subject: the base member

Relation: REFERENCE

->An oral appliance according to any one of claims 1 to 3

Relation: REFINEMENT

->the base member consists essentially of a rigid [...]

Example 4.35 Independent Claim After Merging

Subject: An oral appliance

Relation: DESCRIPTION

->for placing in a mouth of a user

Relation: COMPOSITION

->a base member

Relation: COMP_PART_DESCRIPTION

->having a generally U-shaped form corresponding
to the outline of a jaw of a user [...]

Relation: REFINEMENT

->the base member consists essentially of a rigid [...]

Concept Extraction

The words annotated with a *New-Concept* annotation are extracted from each part in order to create an inverted index. A hashtable is used where the extracted terms are used as keys. Since each concept in a claim should be introduced only once it would be sufficient to store only one claim part for each new-concept. Nevertheless in order to keep the data structure more generic a list of parts can be attached to each extracted new-concept. In addition to creating this concept index the extracted terms are also stored in a list in the tree node corresponding to the part from which they were extracted. For ref-concepts it is sufficient to have the extracted phrases available as a list in the corresponding nodes, thus no index is created for these phrases.

Claim Merging

After the concepts are extracted a recursive procedure is used to find a node to which the refinement-parts extracted from dependent claims can be attached to. The pseudo code of the three most important functions of the algorithm is shown in Listings 4.7, 4.8 and 4.9.

Listing 4.7 shows the method which is initially called for each independent apparatus claim. The method starts the claim merging process by calling the *merge* method with the independent claim as an argument (line 2). The *merge* method returns a list of claims for which no node was found to which their refinement-part can be attached to. Since these claims should not be lost their refinement-parts are directly attached to the claim-subject node of the independent claim (lines 3-5). In Listing 4.9 the core function of the algorithm is shown. The function is recursively called for each claim directly or indirectly related to the independent claim (line 13). Since the same dependent claim can reference several previous claims the algorithm takes care that each claim is processed only once by terminating the recursion if a claim has already been merged (lines 5-9). The algorithm first tries to attach the refinement-part of each dependent claim to a node in its direct parent claim (lines 28-37). If the subject of the dependent claim is an empty string or if the refinement-part can not be attached to any node, the claim is added to a list of unattached claims. This list is returned to the caller of the method (line 39). Duplicate claims are filtered from the list (lines 15-19). In lines 21-26 the algorithm tries to attach the refinement-part of each of these unattached claims to a node in the currently processed claim which is one of its indirect parents. The actual matching procedure is shown in Listing 4.8. The method iterates over all new-concepts in the concept index

of the parent claim and attaches the refinement-part of the dependent claim to node in the tree structure of the parent claim with the best matching new-concept. If the concept index contains several nodes for a given concept, the refinement-part is attached only to the first node since attaching the same part to several nodes might decrease readability. A Levenshtein distance is used as similarity measure. A similarity value (a value between 0 and 1) is computed for the dependent-claim-subject and each new-concept in the concept index. The minimum required similarity value for the claim-subject and the new-concept is set to 0.7. This is a tradeoff between allowing only perfect matches and reducing the number of incorrect attachments.

```

1  FUNCTION: void mergeIndependentClaim(Claim)
2      Unattached-Claims := CALL merge(Claim)
3      FOR EACH Unattached-Claim IN Unattached-Claims
4          ATTACH Unattached-Claim.RefinementPart to Claim.
              ClaimSubject
5      END FOR
6  END FUNCTION

```

Listing 4.7: Function: Merge Independent Claim

```

1  FUNCTION: boolean attachClaim(Claim, Related-Claim){

3      MaxSimilarity := -1

5      FOR EACH New-Concept in Claim.ConceptIndex
6          Similarity := CALL sim(New-Concept, Related-Claim.
              ClaimSubject)
7          IF Similarity > MaxSimilarity
8              MaxSimilarity := Similarity
9              Found-Concept := New-Concept
10         END IF
11     END FOR

13     IF MaxSimilarity > 0.7
14         Node := GET first node FROM Claim.ConceptIndex[Found-
              Concept]
15         ATTACH Related-Claim.RefinementPart to Node
16         RETURN(TRUE)
17     ELSE
18         RETURN(FALSE)
19     END IF
20 END FUNCTION

```

Listing 4.8: Function: Attach Claim

```

1  FUNCTION: LIST merge(Claim)

3      LIST Unattached-Claims := []

5      IF Claim.Merged = TRUE THEN
6          RETURN(Unattached-Claims)
7      END IF

9      Claim.Merged := TRUE

11     FOR EACH Related-Claim IN Claim.RelatedClaims

13         Unattached-Related-Claims :=CALL merge(Related-Claim)

15         FOR EACH Unattached-Related-Claim IN Unattached-
16             Related-Claims
17             IF NOT Unattached-Claims CONTAINS Unattached-
18                 Related-Claim
19                 ADD Unattached-Related-Claim TO Unattached-Claims
20             END IF
21         END FOR

22     FOR EACH Unattached-Claim IN Unattached-Claims
23         Attached :=CALL attachClaim(Claim,Unattached-Claim)
24         IF Attached = TRUE
25             REMOVE Unattached-Claim FROM Unattached-Claims
26         END IF
27     END FOR

28     IF Related-Claim.ClaimSubject is Empty
29         ADD Related-Claim to Unattached-Claims
30     ELSE
31         Attached :=CALL attachClaim(Claim, Related-Claim)
32         IF Attached = TRUE
33             IF Unattached-Claims NOT CONTAINS Related-Claim
34                 ADD Related-Claim to Unattached-Claims
35             END IF
36         END IF
37     END IF

39     RETURN(Unattached-Claims)

41     END FOR

43     END FUNCTION

```

Listing 4.9: Function: Merge

Reattachment of Claim Parts

In some cases nested sentences or characterized-parts extracted from independent claims are not attached to the node where the element they refine was introduced. Thus a similar procedure as for attaching the refinement-parts extracted from dependent claims is used for reattaching these parts. The first ref-concept found in the nested sentence or characterized-part is used to find nodes in the tree structure where the parts may be attached to. For this purpose a similarity measure is computed for the selected ref-concept and each new-concept in the concept index of the independent claim. The part is reattached to the node with the best matching new-concept provided that the Levenshtein similarity value for the two concepts is larger than 0.7. Otherwise the part remains attached to its original parent. In Example 4.36 the tree structure of an independent claim is shown before the reattachment procedure has been executed. The nested sentence “the aspiration piece being connected to the entrance passage” is incorrectly attached to the composition-part “a first end connection”. The reattachment algorithm finds the correct parent of the nested sentence by matching the first ref-concept (“the aspiration unit”) in the nested sentence with all new-concepts in the concept index of the independent claim. It reattaches the nested sentence to the composition-part-description where the refined concept was introduced. Example 4.37 shows the tree after the part has been reattached.

Example 4.36 EP1457216-A2: Tree Before Execution of Reattachment Procedure

Subject: A filtering device

Relation: COMPOSITION

->a first end connection

Relation: COMP_PART_DESCRIPTION

->for coupling an aspiration piece to one end of the
body

Relation: NESTED_SENTENCE

->the aspiration piece being connected to the entrance
passage

Example 4.37 EP1457216-A2: Tree Before Execution of Reattachment Procedure

Subject: A filtering device

Relation: COMPOSITION

->a first end connection

Relation: COMP_PART_DESCRIPTION

->for coupling an aspiration piece to one end of the
body

Relation: NESTED_SENTENCE_REATTACH

->the aspiration piece being connected to the
entrance passage

Summary

This chapter introduced the rule-based decomposition method developed in this work and described the data sets used for creating and evaluating the method. The decomposition method consists of a number of sequentially applied steps. In the first step the claims in English language are extracted from the patent documents. Several preprocessing steps for removing references to images and normalizing spelling of important keywords are applied to the original claim texts. The claims are then classified into dependent and independent claims. For independent claims it is further differentiated between physical entity claims, method claims and use claims. The claims are decomposed by looking for extractable grammatical patterns like nested sentences or enumerations of elements of an invention. The rules for identifying these patterns are based on linguistic analysis of the claims and vary depending on the claim category. The focus of the method is the decomposition of physical entity claims. For each claim a tree structure is created from the extracted parts. In the last step of the method the tree structures from dependent and independent claims are merged by attaching refinements defined in dependent claims directly to the part where the refined concept was introduced.

Chapter 5

Evaluation

Abstract

This chapter provides an evaluation of the developed method. In Section 5.1 it is evaluated how the developed method can be used to reduce length and complexity of independent claims. Section 5.2 provides a quality estimation for the developed rule set by manually evaluating the decomposition trees of physical entity claims and method claims. In Section 5.3 the procedure for merging the tree structures of decomposed dependent and independent physical entity claims is evaluated.

5.1 Independent Claims: Length and Complexity Reduction

In this section it is evaluated how the method developed in this work reduces the length and complexity of independent claims. For evaluating length reduction the average length of the original independent claims is compared with the average length of parts extracted from these claims. The complexity reduction is measured by applying the Stanford Parser to the extracted parts and comparing the percentage of successful parses with the percentage of successful parses of the original claims.

5.1.1 Length Reduction

Table 5.1 shows the number of extracted parts and the average number of words per part for the Analyzed Set and the Evaluation Set and compares them to the average claim length of the unparsed claims. The application of

the extraction algorithm shows very promising results in terms of length reduction of independent claims. For the Analyzed Set the average part length is reduced by about 85% compared to the original claim length. For the Evaluation Set a reduction of about 87% is achieved. The results incorporate all extracted claim parts except the claim-subject since it normally consists of only about three words and would therefore distort the average number of words per part and the average number of successful parses.

The good performance on the Evaluation set indicates that the rules are generic enough to achieve a high reduction of complexity for all patents from the IPC category A61C. It also indicates that the decomposition algorithm can not only be applied to European patents but can also handle the structurally slightly different US patents.

Data set	# Parts	Avg. claim length	Avg. part length
Analyzed Set	1,012	127.81	18.95
Evaluation Set	100,291	132.33	16.95

Table 5.1: Length Reduction: Independent Claims

Table 5.2 compares, for both data sets, the average length of parts extracted from physical entity claims with the average length of parts extracted from claims belonging to the other two categories. The figures show that the average length of physical entity claim parts is less than half of the average length of method and use claim parts. This reflects the fact that the decomposition rule set for physical entity claims is much larger than the one for method claims and shows the positive results of decomposing extracted claim parts into smaller sub-parts.

Data set	Claim category	# Parts	Avg. part length
Analyzed Set	Physical Entity claims	859	15.90
	Method and Use claims	153	36.06
Evaluation Set	Physical Entity claims	85,757	15.16
	Method and Use claims	14,534	27.54

Table 5.2: Length Reduction Comparison for Claim Categories

5.1.2 Complexity Reduction

The achieved complexity reduction can be estimated from the number of successful parses using the Stanford Parser. Table 5.3 shows the success rate of the parser applied to the parts extracted from the Analyzed Set and the

Evaluation Set with the same JVM heap size settings used for parsing the original non decomposed claims (cf. Tables 4.3 and 4.4).

Data set	JVM max. heap size	Successful parses	Failed parses	% of successful parses
Analyzed Set	1000MB	1,010	2	99.80%
	500MB	1,003	9	99.11%
Evaluation set	1000MB	100,140	151	99.85%
	500MB	99,793	498	99.50%

Table 5.3: Stanford Parser Success Rate: Extracted Parts

In Figure 5.1 the percentage of successful parses with the parser applied to the original claims and the extracted parts are compared. The first two letters in the X-axis indicate the patent collection (AS = Analyzed Set, ES = Evaluation Set). The following number shows the JVM maximum heap size settings (500MB or 1000MB). The percentage of successful parses is plotted on the Y-axis.

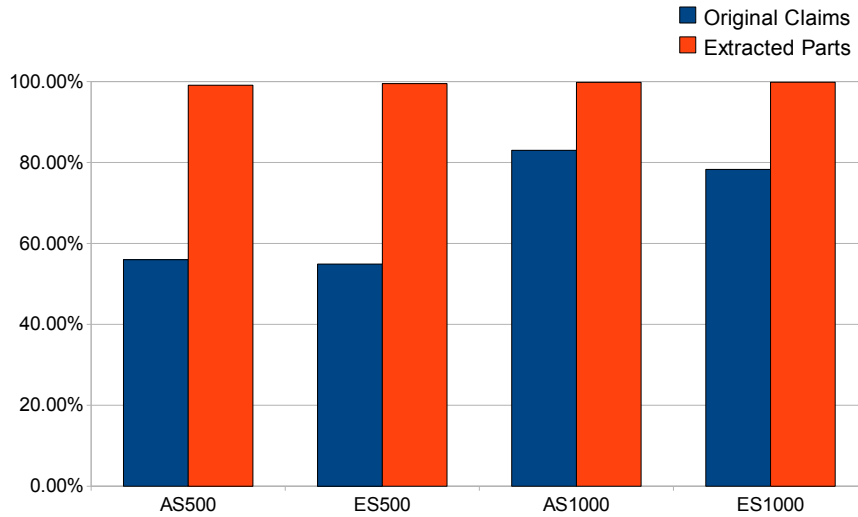


Figure 5.1: Stanford Parser Success Rate: Comparison between Extracted Parts and Original Claims

The comparison shows that the Stanford Parser performs significantly better on the extracted parts than on the original claims. For the Analyzed

Set we were able to raise the percentage of successful parses from 83.01% to 99.80% with a JVM maximum heap size of 1000MB. Due to the significantly reduced average number of words in the extracted parts compared to the original claims, the improvement is even higher with a JVM maximum heap size of 500MB. The percentage of successful parses increased from 55.97% to 99.11%. Only two extracted parts can not be successfully parsed with 1000MB of maximum heap size. Both parts are extracted from method claims for which fewer extraction rules are available than for physical entity claims. The parts contain a length of 280 and 201 words. With 500MB of maximum heap size the parser fails to parse seven additional parts having an average length of 137.8 words. All these additional parts are composition-parts extracted from method claims which are not decomposed further into smaller parts as opposed to composition-parts from physical entity claims.

The improvement for the Evaluation set is slightly higher rising from 78.30% to 99.85% for a JVM maximum heap size of 1000MB and from 54.90% to 99.50% for a JVM maximum heap size of 500MB. This indicates a correlation between the reduction of length and the reduction of complexity.

5.2 Quality Estimation of Independent Claim Decomposition

In order to provide an estimation of the quality of the rule sets 15 physical entity claims selected from 15 different patents and 10 method claims selected from 10 different patents, were manually analyzed and checked for correctness. Due to their small number in both data sets use claims were excluded from the evaluation. Since no gold standard is available this evaluation was done in an informal way by manually classifying the claims as “correct/mostly correct”, “partly correct” and “incorrect/insufficiently decomposed”. This section describes obvious decomposition errors, flaws and weaknesses of the developed rules and possible improvements of them.

5.2.1 Physical Entity Claims

Overall Quality Estimation The overall quality estimation of the decomposition rules for physical entity claims is very promising in terms of accuracy and coverage. Most of the evaluated claims are either decomposed correctly or with minor errors. Only very few claims were found which are classified as physical entity claims but are structurally too different to be handled properly by the rules. The evaluation results are shown in Table 5.4. From the 15 analyzed claims 9 are decomposed correctly or almost correctly, 2 are

	Count	Percentage
Correct	9	60.00%
Partially correct	2	13.33%
Incorrect	4	26.67%

Table 5.4: Quality Estimation: Physical Entity Claims

considered partially correct and 4 are classified as incorrect or insufficiently decomposed.

Example 5.1 shows a claim having a structure which can be decomposed very accurately with the developed decomposition rules. It contains a Composition-Pattern from which several composition-parts are extracted correctly which themselves can be decomposed further into smaller parts. In addition to that a Characterized-Pattern and several Nested-Sentence-Patterns are identified by the rules. The result is a very fine-grained decomposition which can be seen in Example 5.2.

Example 5.1 EP1558168-B1: Original Claim

An orthodontic appliance for a molar tooth comprising: a base for connecting the appliance to a molar tooth; a body extending from the base, the body having a mesial side portion and a distal side portion; a mesial archwire guide connected to the mesial side portion of the body; a distal archwire guide connected to the distal side portion of the body; an archwire slot extending across the mesial archwire guide and the distal archwire guide in a generally mesial-distal direction; and a latch for releasably retaining an archwire in the archwire slot , the latch being movable from a slot-open position for admitting the archwire in the archwire slot and to a slot-closed position for retaining the archwire in the archwire slot , wherein the appliance lacks tieings, wherein the distal archwire guide is spaced from the mesial archwire guide to present a channel that extends in a generally occlusal-gingival direction, wherein the latch is remote from the channel characterized in that the channel and the archwire slot each have a lingual side, the lingual side of the channel being spaced in a lingual direction from the lingual side of the archwire slot .

Example 5.3 on the other hand shows a claim which is classified as physical entity claim but which can not be decomposed into sufficiently small parts by the developed rules. In Example 5.4 the decomposition of the claim is shown. Besides the claim-subject only three parts are extracted. Except of the Characterized-Pattern from which the two composition-parts are extracted no other patterns can be found. Therefore the parts remain too long and complex having an average length of 45.3 words.

Example 5.2 EP1558168-B1: Correctly Decomposed Claim

Subject: An orthodontic appliance

Relation: CHARACTERIZED

->The channel and the archwire slot each have [...]

Relation: DESCRIPTION

->for a molar tooth

Relation: COMPOSITION

->a base

Relation: COMP_PART_DESCRIPTION

->for connecting the appliance to a molar tooth

->a body

Relation: COMP_PART_DESCRIPTION

->extending from the base

Relation: NESTED_SENTENCE

->the body having a mesial side portion and a
distal side portion

->a mesial archwire guide

Relation: COMP_PART_DESCRIPTION

->connected to the mesial side portion of the body

->a distal archwire guide

Relation: COMP_PART_DESCRIPTION

->connected to the distal side portion of the body

->an archwire slot

Relation: COMP_PART_DESCRIPTION

->extending across the mesial archwire guide and
the distal archwire guide in a generally
mesial-distal direction

->a latch

Relation: COMP_PART_DESCRIPTION

->for releasably retaining an archwire in
the archwire slot

Relation: NESTED_SENTENCE

->the latch being movable from a slot-open position
for [...]

->the appliance lacks tieings

->the distal archwire guide is spaced from [...]

->the latch is remote from the channel

Example 5.3 EP0171002-B1: Original Claim

An impression tray for use with dental impression material which impression tray is formed from transparent relatively rigid material and has a recess adapted to hold a predetermined amount of the impression material for the impression of dental anatomy thereinto, characterized in that for use with dental impression material capable of being polymerized by exposure to visible actinic light the tray has light-receiving means being integral with the tray and being adapted to receive and transmit light to the tray ; and the tray also comprises a light-reflecting means arranged at the exterior surfaces of the tray, preventing the passage of light rays from the exterior of the tray to the interior thereof and reflecting light applied through the light-receiving means into the dental impression material within the recess to effect polymerization thereof to a degree that it has a permanent elastomeric form.

Example 5.4 EP0171002-B1: Insufficiently Decomposed Claim

Subject: An impression tray

Relation: DESCRIPTION

->for use with dental impression material which impression tray is formed from transparent relatively rigid material and has a [...]

Relation: CHARACTERIZED_COMPOSITION

->For use with dental impression material capable of being polymerized by exposure to visible actinic light the tray has light-receiving means being [...]

->the tray also comprises a light-reflecting means arranged at the exterior surfaces of the tray, preventing [...]

Quality estimation for individual patterns In order to provide a better quality estimation, the rule set for each of the following patterns is analyzed independently.

- Claim-Subject Extraction
- Description-Pattern
- Composition-Pattern
- Characterized-Part Decomposition
- Nested-Sentence-Pattern

Claim-Subject Extraction For all evaluated claims it is correct to extract the first noun phrase as claim-subject. During the evaluation no case was found where the first noun phrase in the claim does not represent the subject and thus a different part should be used as root of the tree structure.

Nevertheless the claim-subject is not always extracted correctly. The correctness of the claim-subject extraction rule is highly dependent on the correctness of the NP-chunker which in turn is dependent on the accuracy of the POS-tagger. This means that if the noun phrase is not correctly recognized by the chunker, the claim-subject is not extracted correctly. This is shown in Example 5.5 where only the words “A substantially cylindrical dental” are extracted as claim-subject instead of the phrase “A substantially cylindrical dental implant anchor”.

Example 5.5 EP0412845-B1: Incorrectly Extracted Claim-Subject

A substantially cylindrical dental implant anchor comprising [...]

Subject: A substantially cylindrical dental

Relation: DESCRIPTION

->implant anchor

The words “implant” and “anchor” are incorrectly tagged as verbs. Therefore the noun phrase is not identified correctly. If the words were tagged as nouns, the noun chunk “A substantially cylindrical dental implant anchor” could be correctly identified. This is illustrated in Example 5.6.

Example 5.6 EP0412845-B1: Illustration of Incorrect Noun Phrase Chunking

Incorrectly tagged phrase:

Correctly tagged phrase:

Description-Pattern The length of the description-part largely depends on whether Composition, Nested-Sentence or Characterized-Patterns are identified correctly. The obvious reason for this is that everything between the claim-subject and the first occurrence of one of these patterns forms the description-part. In most cases the description-part is short and precise such as shown in Example 5.7.

Example 5.7 EP1543792-A1: Short and Precise Description-Part

A prophy chip, mounted on the top of a dental rotary instrument for cleaning, polishing, and burnishing teeth, comprising [...]

Subject: A prophy chip

Relation: DESCRIPTION

->mounted on the top of a dental rotary instrument for
cleaning, polishing, and burnishing teeth

Since the description-parts are currently not decomposed further the rule set could be improved by extracting Composition-Patterns. A description-part from which a Composition-Pattern could be extracted is shown in Example 5.8. A modification of the Composition-Pattern rules would be necessary with respect to the keywords introducing the pattern, where the word “having” has to be matched instead of the keywords “comprising”, “comprises” or “including”. For decomposing the claims shown in Example 5.8 the rules for extracting the composition-parts could be left unchanged. The result would

be the extraction of four composition-parts: “a labial surface”, “a pair of proximal surfaces”, “an incisal surface” and “a lingual surface”.

Example 5.8 EP0472656-B1: Description with Unextracted Composition-Pattern

Apparatus for use during restoration of a tooth having a labial surface, a pair of proximal surfaces, an incisal surface, and a lingual surface, the apparatus comprising [...]

Subject: Apparatus

Relation: DESCRIPTION

->for use during restoration of a tooth having a labial surface, a pair of proximal surfaces, an incisal surface, and a lingual surface

In cases where no pattern which ends the description-part can be found, the description-part remains long and complex. An example for this case is Example 5.4 which has already been examined in the previous paragraph.

Composition-Pattern The Composition-Pattern rule set is the one which contributes most to length and complexity reduction of claims since this pattern occurs in almost all independent physical entity claims. Examples 5.9 and 5.10 show a rather long claim and its decomposition. A high number of composition-parts and their descriptions are extracted correctly.

Example 5.9 EP1558168-B1: Original Claim

An orthodontic appliance for a molar tooth comprising: a base for connecting the appliance to a molar tooth; a body extending from the base, the body having a mesial side portion and a distal side portion; a mesial archwire guide connected to the mesial side portion of the body; a distal archwire guide connected to the distal side portion of the body; an archwire slot extending across the mesial archwire guide and the distal archwire guide in a generally mesial-distal direction; and a latch for releasably retaining an archwire in the archwire slot [...]

The quality of the extracted Composition-Patterns is high in the sense that no Composition-Patterns in the evaluated claims were missed by the rules. In addition most composition-parts are extracted correctly. Nevertheless there are a few cases which the decomposition rules fail to handle correctly. The most common errors are examined in the following paragraphs.

Example 5.10 EP1558168-B1: Correctly Decomposed Composition-Pattern

Subject: An orthodontic appliance

Relation: DESCRIPTION

->for a molar tooth

Relation: COMPOSITION

->a base

Relation: COMP_PART_DESCRIPTION

->for connecting the appliance to a molar tooth

->a body

Relation: COMP_PART_DESCRIPTION

->extending from the base

Relation: NESTED_SENTENCE

->the body having a mesial side portion
and a distal side portion

->a mesial archwire guide

Relation: COMP_PART_DESCRIPTION

->connected to the mesial side portion of the body

->a distal archwire guide

Relation: COMP_PART_DESCRIPTION

->connected to the distal side portion of the body

->an archwire slot

Relation: COMP_PART_DESCRIPTION

->extending across the mesial archwire guide and the
distal archwire guide in a generally mesial-distal
direction

->a latch

Relation: COMP_PART_DESCRIPTION

->for releasably retaining an archwire in the
archwire slot

Incorrect Split of Composition-Part Due to the characteristics of the rule for separating the composition-part and its description it occurs frequently that the composition-part is split incorrectly and that parts of it are moved to the description. This is illustrated in Example 5.11 where the composition-part consists only of the word “light” instead of the terms “light transmitting means”.

Example 5.11 EP1558168-B1: Incorrectly Split Composition-Part

An irradiation device for [...] comprising [...] light transmitting means having a tip extending [...]

Subject: An irradiation device

Relation: COMPOSITION

->light

Relation: COMP_PART_DESCRIPTION

->transmitting means having a tip extending beyond the forward end of the housing

Unextracted Composition-Parts Although the rules for extracting the composition-parts from a Composition-Pattern work well in most cases there are claims where certain patterns are used which the rules can not identify correctly. In Example 5.12 the two highlighted elements of the invention are not introduced with an indefinite article and are thus not recognized as separate composition-part.

Example 5.12 EP0453493-B1: Unextracted Composition-Part

A syringe for washing teeth root canals comprising: a tubular guide to [...]; **helical spring to [...]** ; **two-way valve body connected to [...]** ; an extension element , [...]

A similar case is shown in Example 5.13 where the hyphen (“-”) causes the rule to fail in extracting the composition-part starting with “a head portion”.

Example 5.13 EP1576935-A1: Unextracted Composition-Part

An endosseous dental implant comprising: an anchoring portion [...]; – a head portion disposed at a first upper end of the anchoring portion , [...]

Nested Composition-Patterns Other structures which can result in erroneous decompositions of claims are nested Composition-Patterns which are for example used to describe one part of an invention in detail. In most cases the grammatical structures used for this purpose are different to normal Composition-Patterns and do not result in an erroneous decomposition. Such a case is shown in Example 5.14. In the nested Composition-Pattern the composition-parts “a cup yoke [...]” and “a cylindrical permanent magnet” are not separated by a semicolon or a comma. Therefore they do not interfere with the extracted Composition-Pattern. Since no rules are currently available for decomposing such nested Composition-Patterns the part is simply attached as a composition-part-description.

Example 5.14 EP1457168-A1: Non-Interfering Nested Composition-Pattern

A dental magnetic attachment comprising a keeper, and a magnet structure comprising a cup yoke formed of a soft magnetic material and a cylindrical permanent magnet [...].

Subject: A dental magnetic attachment

Relation: COMPOSITION

->a keeper

->a magnet structure

Relation: COMP_PART_DESCRIPTION

->comprising a cup yoke formed of a soft magnetic material and a cylindrical permanent magnet

Nevertheless there are cases where the nested Composition-Patterns interfere with the decomposition of the parent Composition-Pattern resulting in an erroneous extraction and attachment of composition-parts. In the claim shown in Example 5.15 the elements of the nested Composition-Pattern are separated by commas and thus extracted by the decomposition rule for the parent Composition-Pattern. The result is an incorrect attachment of these parts to the claim-subject instead of to the composition-part “a total of four teeth”.

Characterized-Part Decomposition The Characterized-Pattern itself is recognized correctly for all examined physical entity claims. The quality estimation therefore focuses on the more complex task, which is the decomposition of extracted characterized-parts into smaller sub-parts. The analysis shows that many extracted characterized-parts are rather long and complex.

Example 5.15 EP1621157-A1: Incorrectly Decomposed Nested Composition-Pattern

A dental prosthesis [...], comprising a total of four teeth including a first premolar, a second premolar, a first molar and a second molar in a maxilla or [...]

Subject: A dental prosthesis

Relation: COMPOSITION

->a total of four teeth

Relation: COMP_PART_DESCRIPTION

->including a first premolar

->a second premolar

->a first molar and a second molar in a maxilla or [...]

Example 5.16 EP0231166-B1: Claim with Complex Characterized-Part

[...], characterized in that the cavities are undercut cavities and extend at either side of a slot having a minimum width of 30 m, through which the cavity is accessible from the outside of the element, **and that** the slot is located on the element so as to extend substantially parallel to the surface of the skin when the element is implanted in the intended position thereof, the cavity having a minimum depth of 30 m

Example 5.17 EP0231166-B1: Improved Decomposition of Characterized-Part

[...]

Relation: CHARACTERIZED

->the cavities are undercut cavities and extend at either side of a slot having a minimum width of 30 m, through which the cavity is accessible from the outside of the element

->the slot is located on the element so as to extend substantially parallel to the surface of the skin when the element is implanted in the intended position thereof

Relation: NESTED_SENTENCE

->the cavity having a minimum depth of 30 m

Example 5.16 shows an extracted characterized-part that still consists of 72 words and contains patterns which would lend themselves to further decomposition. With an improved rule set it would be possible to split the characterized-part at the highlighted phrase. In addition to that the decomposition could be improved by applying the already existing Nested-Sentence-Pattern rule. This would result in the improved decomposition illustrated in Example 5.17.

The rules for extracting Composition-Patterns from characterized-parts are very difficult to evaluate since only very few characterized-parts contain Composition-Patterns introduced by the keywords “including”, “comprising” or “comprises”. One case where this pattern occurs and where it is correctly decomposed is shown in Example 5.18. From the otherwise long and complex characterized-part (88 words) a Composition-Pattern with two composition-parts is extracted correctly leading to a high reduction of the average part length. In cases where the Composition-Pattern can not be decomposed correctly the errors result from the same grammatical structures already described in the evaluation of the decomposition rules for Composition-Patterns.

Example 5.18 EP1384449-B1: Characterized-Pattern with Composition-Pattern Correctly Extracted

An apical foramen locator comprising [...] characterized in that it further comprises: at least one impedance map including apical foramen location data corresponding to a combination of a first voltage index and a second voltage index wherein the apical foramen location data is generated from reference teeth; and a processing component operable to derive the first and second voltage indices from the voltages sensed by the impedance-sensing circuit and to select from the impedance map apical foramen location data that corresponds to the first and second voltage indices.

Subject: An apical foramen locator

Relation: CHARACTERIZED

->It further

Relation: COMPOSITION

->at least one impedance map

Relation: COMP_PART_DESCRIPTION

->including apical foramen location data [...]

->a processing component

Relation: COMP_PART_DESCRIPTION

->operable to derive the first and second voltage [...]

In most cases, however, the nested Composition-Pattern is simply a list of refinements separated by semicolons such as shown in Example 5.19. In all evaluated cases this pattern was extracted correctly from the characterized-part.

Example 5.19 EP0471680-B1: Characterized-Part with List of Refinements Correctly Extracted

A rack for instruments [...] characterized in that walls separating the compartments of the drum are sealed relative to the outer casing ; and that the outer casing is provided with means for sterilizing the instruments when they are not being used.

Subject: A rack

Relation: CHARACTERIZED_COMPOSITION

->walls separating the compartments of the drum are
sealed relative to the outer casing
->that the outer casing is provided with means for
sterilizing the instruments when they are not being used

Nested-Sentence-Pattern The quality of the extracted Nested-Sentence-Patterns is very high and contributes greatly to the achieved length and complexity reduction. In Examples 5.20 and 5.21 a claim which consists of almost only Nested-Sentence-Patterns and its decomposition are shown. In the evaluated claims no case was found where a nested sentence was incorrectly extracted.

Example 5.20 EP0028529-B2: Original Claim

A scaler tip having an operative end and an end adapted to be secured to a hand-held vibratory scaler , the operative end terminating in a curved free end , the operative end having a non-abrasive working portion in which [...], the curved free end lying in a plane passing through [...] , the plane also passing through [...].

Example 5.21 EP0028529-B2: Claim composed of almost only Nested-Sentence-Patterns

Subject: A scaler tip

Relation: DESCRIPTION

->having an operative end and an end adapted to be secured
to a hand-held vibratory scaler

Relation: NESTED_SENTENCE

->the operative end terminating in a curved free end
->the operative end having a non-abrasive working portion
in which [...]
->the curved free end lying in a plane passing through [...]
->the operative end being substantially symmetrical
about the plane passing through [...]
->the plane also passing through [...]

5.2.2 Method Claims

Table 5.5 shows the evaluation results for the 10 analyzed claims. The figures show that 4 claims are decomposed correctly, 2 are partially correct and 4 are insufficiently or incorrectly decomposed. The detailed evaluation shows that the performance of the developed decomposition rules varies greatly depending on the structure of the claims. Method claims which consist of an enumeration of steps, wherein each step starts with a verb in gerund form, are decomposed correctly. Some claims on the other hand also provide a description of materials or apparatuses used for carrying out the method or enumerate steps in a form that can not be handled correctly by the rules.

	Count	Percentage
Correct	4	40.00%
Partially correct	2	20.00%
Incorrect	4	40.00%

Table 5.5: Quality Estimation: Method Claims

Example 5.22 shows a correctly decomposed method claim. It contains a Composition-Pattern from which several composition-parts are correctly extracted.

There are many claims which are decomposed correctly but where the extracted parts are still very long and complex. This can occur for extracted composition-parts as well as characterized-parts.

Example 5.22 EP0154137-B1: Correctly Decomposed Method Claim

A method of preparing a dental restoration comprising preparing an opaque dental mount having a specific chroma on the Munsell chroma scale and a specific Munsell hue; placing a crown, which is substantially hue and chroma free and translucent throughout, on the dental mount; applying shader to the crown and viewing the opaque dental mount while applying the shader; and removing the shaded crown from the opaque dental mount.

Subject: A method

Relation: DESCRIPTION

->preparing a dental restoration

Relation: COMPOSITION

->preparing an opaque dental mount having a specific chroma
on the Munsell chroma scale and a specific Munsell hue
->placing a crown, which is substantially hue and chroma
free and translucent throughout, on the dental mount
->applying shader to the crown and viewing
the opaque dental mount while applying the shader
->removing the shaded crown from the opaque dental mount

Example 5.23 shows a method claim where a Composition-Pattern consisting of two composition-parts is extracted correctly but where the second part still has a length of 48 words. In order to improve the performance of method claim decomposition, rules have to be developed for splitting these parts into smaller units.

There are certain cases where the decomposition rules are not suited for the structure of the claim. One case is illustrated in Example 5.24. The claim does not enumerate the steps of the method in a Composition-Pattern which can be extracted by the developed rule set. It focuses on the description of the materials which are used and describes the steps for carrying out the method in the characterized-part in a form which can not be handled by the existing rules.

Example 5.23 EP1410768-B1: Insufficiently Decomposed Method Claim

A method for [...], comprising the steps of: providing a plurality of light sources, wherein each of the plurality of light sources produces an incident light beam; and integrating each of the incident light beams into an output light beam, the output light beam having an output power intensity distribution; wherein a first one of the light sources has a first characteristic wavelength and a second one of the light sources has a second characteristic wavelength.

Subject: A method

Relation: COMPOSITION

->providing a plurality of light sources, wherein each of the plurality of light sources produces an incident light beam
 ->integrating each of the incident light beams into an output light beam, the output light beam having an output power intensity distribution; wherein a first one of the light sources has a first characteristic wavelength and a second one of the light sources has a second characteristic wavelength

Example 5.24 EP0415508-B1: Incorrectly Decomposed Method Claim

A method for continuously hardening an object composed of visible light-curing resin, wherein the resin contains a polymerization initiator which is photosensitive to a wavelength within the visible light range, for example a camphor quinone, wherein the object is irradiated with visible light comprising a wavelength component and an illuminance component whilst being continuously advanced to harden the resin of which it is formed, characterized in that the object to be irradiated is advanced through a plurality of visible light irradiation stations, the visible light in each station being derived from a light source independently selected from a halogen lamp, a xenon lamp, a short arc lamp and a fluorescent lamp, and the visible light in each station being irradiated without changing the wavelength component and changing only the illuminance component.

5.3 Claim Merging

From each of the data sets 10 patents containing a physical entity claim were selected randomly and evaluated manually in terms of correct attachments, incorrect attachments and the number of parts for which no attachment was found. For the parts which could not be attached it is differentiated between parts for which no claim-subject was found and those part which could not be attached although a claim-subject was identified by the rules. For the dependent claims for which no subject could be found it is analyzed whether the claim-subject does not exist or whether it was not identified by the decomposition rules.

	Total Number	Percentage
Attached claim references	81	96.43%
Missing claim references	3	3.57%
Total number of dependent claims	84	100%

Table 5.6: Resolved Claim References: Analyzed Set

	Total Number	Percentage
Attached claim references	77	100%
Missing claim references	0	0%
Total number of dependent claims	77	100%

Table 5.7: Resolved Claim References: Evaluation Set

Tables 5.6 and 5.7 show the performance of the rules used for resolving references from dependent claims. The row “Attached claim references” shows for how many dependent claims the reference to their parent was correctly resolved while the row “Missing claim references” shows how many claims could not be attached to the claim they refine. The sum of the numbers in these two rows is shown in the row “Total number of dependent claims”. The figures show that for all independent claims selected from the Evaluation set the dependent claims were attached successfully. In the Analyzed Set the claim reference was not extracted successfully for two dependent claims. The third claim which could not be attached references one of these claims and can therefore not be added to the data structure. For the two other claims the rule used to identify the references is not generic enough. In the first claim shown in Example 5.25 the reference “the previous claims” is not resolved and in the second claim shown in Example 5.26 a range is specified in the format “from 1 to 3” which is not recognized.

Example 5.25 EP1457216-A2: Unresolved Claim Reference

The filtering device according to **the previous claims**, wherein the first connection and the second connection are identical in construction, and wherein only the first connection is provided with a diverter means.

Example 5.26 EP0453493-B1: Unresolved Range

A syringe according to any of the claims **from 1 to 3** wherein the first and second spring means capable of constantly urging the first and second obturator in the closing position of the longitudinal duct and the transversal duct respectively, in the two-way valve body consist of helicoidal springs.

Tables 5.8 and 5.9 provide an overview of the performance of the claim merging process for the Analyzed Set and the Evaluation set. The row “Correct attachments” shows how many parts were attached correctly to the part they refine and the row “Incorrect attachments” shows how many parts were attached erroneously.

	Total Number	Percentage
Correct attachments	33	40.74%
Incorrect attachments	5	6.17%
No attachment found	24	29.63%
No claim-subject/correct	9	11.11%
No claim-subject/incorrect	10	12.35%
Attached claim references	81	100%

Table 5.8: Attachments: Analyzed Set

In the row “No claim-subject/correct” it can be seen how many dependent claims did not have an extractable claim-subject. The row “No claim-subject/incorrect” shows for how many dependent claims a claim-subject existed but was not found by the rules. The figures show that the percentage of parts for which no attachment was found is relatively high for both the Analyzed as well as the Evaluation Set while the percentage of correct attachments and incorrect attachments is relatively low. There are several reasons for this.

One reason is that a *Ref-Concept* in a dependent claim can be provided in a shorter form than the original *New-Concept*. This is shown in Example 5.27 where a concept is introduced as “spaced-apart arms” in an independent claim and referenced with “the arms” in the dependent claim.

	Total Number	Percentage
Correct attachments	36	46.75%
Incorrect attachments	1	1.30%
No attachment found	32	41.56%
No claim-subject/correct	2	2.60%
No claim-subject/incorrect	6	7.79%
Attached claim references	77	100%

Table 5.9: Attachments: Evaluation Set

Example 5.27 US20050172982-A1: Concept Referenced in Short Form

Independent claim:

A dental floss holder comprising a base portion and a pair of **spaced-apart arms** [...]

Dependent claim:

A dental floss holder according to claim 4 each of **the arms** comprises a snap-fit projection

This could be compensated by lowering the threshold of the string similarity measure which would increase the number of correct attachments as well as the number of incorrect attachments.

Another reason is that some dependent-claim-subjects are not extracted correctly due to erroneous POS-tagging. This affects especially the term “means” which is always tagged as a verb. This is shown in the dependent claim shown in Example 5.28 where the term “the light-reflecting” is extracted as the claim-subject instead of the term “the light-reflecting means”. A possible solution would be to create a specific rule for the term “means” in a similar way as it is done for extracting composition-parts.

Example 5.28 EP0171002-B1: Incorrect Identification of Claim-Subject due to Erroneous POS-Tagging

The impression tray according to claim 1 in which the **light-reflecting means** comprises a thin layer of reflective metal.

The third main reason is that the extracted claim-subject is not always the concept which is refined. This is shown in Example 5.29 where the term “the edges” is extracted as claim-subject instead of the words “the cover sheet”.

Example 5.29 EP0171002-B1: Claim-Subject not the Refined Concept

The impression tray according to claim 5 in which **the edges of the cover sheet** are sealed to the rim of the tray by cement capable of permitting the cover sheet to be peeled from the rim of the tray to expose the impression material for use.

This problem is also reflected in the number of dependent claims for which erroneously no claim-subject was found. Most of those claims follow a structure where the concept to which the part should be attached is written at the end of the sentence. Such a case is shown in Example 5.30 where “the wire support” should be extracted as the claim-subject.

Example 5.30 EP0453493-B1: Unidentified Claim-Subject

A teeth straightening bracket according to claim 1 characterized in that engaging fingers on the incisal and gingival side of the wire support are disposed except for the both longitudinal ends of **the wire support**.

Summary

This chapter provided an evaluation of the method described in Chapter 4. It was shown that the developed decomposition rules significantly reduce the length and complexity of independent claims. For the evaluated data sets a length reduction of over 85% was achieved and it was shown that the performance of the Stanford Parser was significantly better on the extracted parts compared to the original claims. The quality of the extracted parts was estimated and possible improvements were shown. The evaluation of physical entity claims showed that most claims were decomposed correctly but that the rules could still be improved to achieve better results. For method claims the analysis has shown that the rules can not handle all occurring structural particularities and that further refinements of the rules are necessary to improve the quality of the results. The evaluation of the claim merging procedure showed good results for claim reference resolution. From 161 analyzed dependent claims only 3 could not be attached to the claim they refine. The performance of attaching refinements extracted from dependent claims to the concept they refine could still be improved as the percentage of correct attachments is only around 40%. Only very few parts were attached incorrectly while for a high percentage of parts the algorithm did not find the concept they refine and which as a result had to be attached directly to the root node of the independent claim.

Chapter 6

Application

6.1 Claim Tree Visualization

After the decomposition is done the created tree structures can be visualized for improving readability of patent claims. For this purpose a Java Swing application is used which allows the user to select single files or an entire folder from which all files are decomposed and then displayed. The user interface consists of two parts. A screenshot of the complete claim browser can be seen in Figure 6.1.

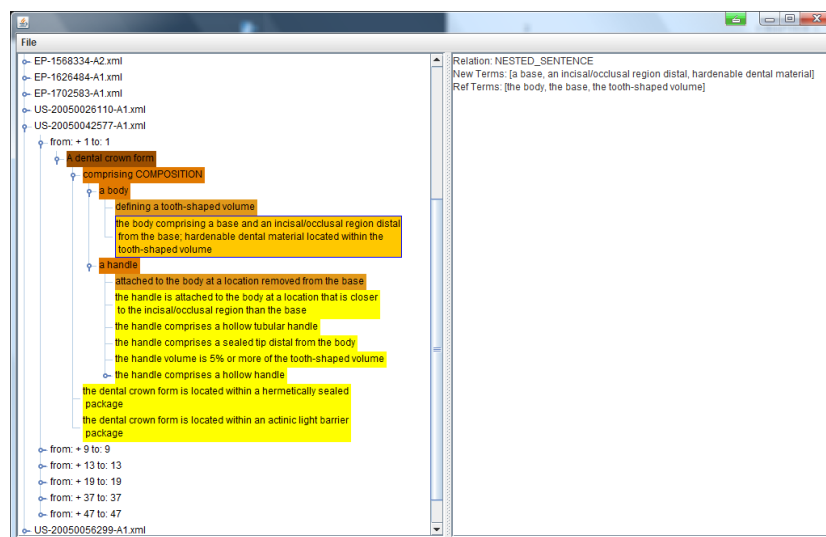


Figure 6.1: Claim Browser: User Interface

On the left side the decomposed claims are displayed as a tree. In this way it is possible to let the user expand and collapse parts of the tree depending

on which part of the claim is currently being examined. Figure 6.2 shows an enlarged screenshot of such a tree. The nodes in the tree have a different color depending on the type of the part they contain. The claim subject is displayed in dark brown, the composition parts in red brown and the composition part descriptions in light brown. The parts displayed in orange are nested sentences and all yellow nodes show refinements extracted from dependent claims.

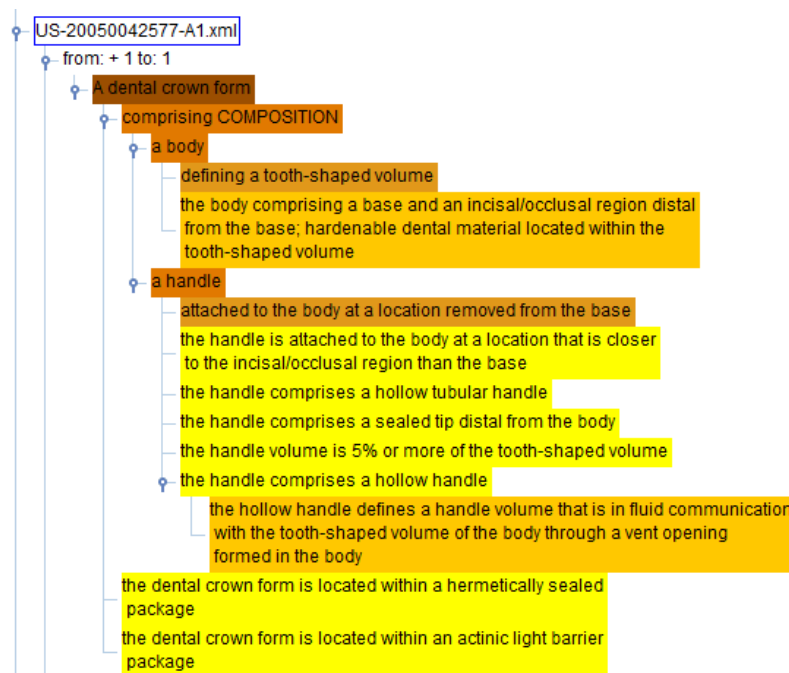


Figure 6.2: Claim Browser: Decomposed Tree

On the right side additional information such as the type of the relation to the parent and new-concepts and ref-concepts are displayed for the currently selected part. For refinements from dependent claims it is shown to which concept the refinement was attached to together with a similarity value. If the claim subject is selected, the decomposition of a claim in textual form is shown. If the claim itself is selected, the right side shows the original claim text before decomposition.

Chapter 7

Conclusions and Future Work

The method developed in the work at hand is a rule-based decomposition method for English-language patent claims. In the first step the claims are extracted from the original patent document. For improving the performance of the decomposition algorithm several preprocessing steps are applied, like normalization of spelling or removing references to images from claims.

Our analysis of European patents has brought us to the conclusion that claims lend themselves very well to rule-based decomposition as they are drafted according to very precise syntactic and semantic rules and thus contain a high number of re-occurring grammatical patterns. The developed rule sets exploit these grammatical patterns in order to identify extractable sub-parts of claims. Almost all of these decomposition rules are based on linguistic analysis, in particular POS-tagging and NP-chunking, and are therefore highly dependent on the quality of the used NLP tool. Due to the invention and patent domain-specific vocabulary used in claims the output of these tools is not always correct. This makes it necessary to adapt the rules accordingly, for taking into account erroneous NP-chunking and incorrect POS-tagging for important keywords.

Due to large structural differences within claims, claim category and claim type specific rule sets are used in the developed method. With respect to claim types it is differentiated between independent claims describing the essential features of an invention and dependent claims which provide refinements to an invention. A claim is assigned to one of three claim categories: method claims, use claims and physical entity claims. For classifying claims into dependent and independent claims as well as for identifying the claim categories simple heuristics based on keyword matching are used.

The patterns identified by the rule sets are extracted and organized in a tree structure in order to retain the information on how the extracted parts are related to each other. The decomposition rules for dependent claims

extract references to refined claims in order to attach dependent claims to the claim they refine. For physical entity claims an algorithm is proposed which merges the tree structures extracted from dependent claims with the tree structure from the claim they refine. This is done by attaching the part describing the refinement directly to the part where the refined element was introduced.

In order to improve the readability of claims and allow an easier examination of decomposed claims, the final tree structure is visualized in a Java application as a tree. Each node of the tree contains an extracted part. This allows the user to collapse and expand parts of the decomposed claims depending on which section is currently being examined.

The developed method shows that rule-based decomposition of patent claims is feasible due to the particular language used for drafting patents. The evaluation shows promising results in terms of reduction of length and complexity of independent claims and shows that the decomposition method eases the application and raises the performance of existing information retrieval and information extraction tools. A quality estimation for the correctness of the extracted parts shows good results for physical entity claims where a high percentage of evaluated claims is decomposed either correctly or with minor errors. While the decomposition rules seem to be detailed enough for physical entity claims, additional work has to be done for method claims as the extracted parts remain very often long and complex. Further analysis has also to be done for dependent method claims for which currently no decomposition rules exist. The procedure for merging dependent and independent claims has to be extended and adapted for method claims. Particularities of dependent method claims will have to be taken into account, as refinements may be provided in different forms than in dependent physical entity claims. Regarding the claim merging procedure for physical entity claims it should be evaluated how the quality of the results changes when different string similarity measures and thresholds are used. It should also be evaluated how the results change when other terms are used for attaching the claim when no attachment can be found for the dependent-claim-subject.

The evaluation on a large data set has shown that the rules created from the analysis of a small data set containing only European patents are generic enough for the IPC category A61C and that they can also be applied to US patents. Since the rule set does not use any domain-specific keywords it is very likely that the rules can also be applied to patents from other IPC categories. To test this hypothesis further evaluation should be done on a data set containing patents from a wider range of IPC categories in order to see how the performance of the rules depends on the domain of the invention.

An important aspect regarding evaluation is to seek intensive cooperation

with researchers from the intellectual property domain for developing precise criteria for measuring the quality and the correctness of the extracted claim parts.

Structural analysis of patent claims has already been done for Japanese patent claims such as in [SOMI03] and has proven to be useful in various fields such as improving readability of patent claims and patent search [TFI04].

To our best knowledge this work is the first approach of decomposing English-language patent claims and can therefore be seen as a starting point for additional work in various fields of patent information retrieval. Besides the visualization of decomposed claims for improving readability as done in this work, the method can be used for tasks such as document retrieval or computing structure-based similarity measures. It can therefore be a contribution to the development of information retrieval methods especially tailored to the patent domain needed by various parties such as patent offices, patent attorneys and inventors.

List of Figures

2.1	IPC Groups A61C	13
5.1	Stanford Parser Success Rate: Comparison between Extracted Parts and Original Claims	88
6.1	Claim Browser: User Interface	109
6.2	Claim Browser: Decomposed Tree	110

List of Tables

4.1	Analyzed Set: Characteristics	43
4.2	Evaluation Set: Characteristics	43
4.3	Stanford Parser Success Rate: Analyzed Set	44
4.4	Stanford Parser Success Rate: Evaluation Set	44
4.5	POS-Tags used in Method Description	51
4.6	Place-Holders used in Method Description	52
4.7	Regular Expression Character Classes	52
4.8	Analyzed Set: Claim Types	59
4.9	Evaluation Set: Claim Types	59
5.1	Length Reduction: Independent Claims	87
5.2	Length Reduction Comparison for Claim Categories	87
5.3	Stanford Parser Success Rate: Extracted Parts	88
5.4	Quality Estimation: Physical Entity Claims	90
5.5	Quality Estimation: Method Claims	102
5.6	Resolved Claim References: Analyzed Set	105
5.7	Resolved Claim References: Evaluation Set	105
5.8	Attachments: Analyzed Set	106
5.9	Attachments: Evaluation Set	107

Bibliography

- [AF05] Makoto Iwayama Atsushi Fujii. Document structure analysis for the NTCIR-5 patent retrieval task. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, Tokyo, Japan*. National Institute of Informatics, 2005.
- [Arc04] Eugenio Archontopoulos. Prior art search tools on the internet and legal status of the results: a european patent office perspective. *World Patent Information*, 26:113 –121, 2004.
- [Atk08] Kristine H. Atkinson. Toward a more rational patent search paradigm. In *Proceedings of the 1st ACM workshop on Patent Information Retrieval (PaIR), Napa Valley, California*, pages 37–40. Association for Computing Machinery, 2008.
- [Bab08] Olga Babina. Nlp-based patent information retrieval. Technical report, South Ural State University, 2008.
- [BCC93] John Broglio, James P. Callan, and W. Bruce Croft. Inquiry system overview. In *Proceedings of a workshop on held at Fredericksburg, Virginia*, pages 47–67. Association for Computational Linguistics, 1993.
- [BJ99] Srinivas Bangalore and Aravind K. Joshi. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25:237–265, 1999.
- [CTA03] Liang Chen, Naoyuki Tokuda, and Hisahiro Adachi. A patent document retrieval system addressing both semantic and syntactic properties. In *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing, Sapporo, Japan*, pages 1–6. Association for Computational Linguistics, 2003.

- [DDF⁺90] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.
- [Emm09] Christiane Emmerich. Comparing first level patent data with value-added patent information: A case study in the pharmaceutical field. *World Patent Information*, 31:117–122, 2009.
- [EPC07] *European Patent Convention*. European Patent Office, 13 edition, July 2007.
- [EPO] Guidelines for examination in the european patent office. <http://www.epo.org/patents/law/legal-texts/guidelines.html>, last visited: 2009-12-08. European Patent Office, Status April 2009.
- [Fal02] Louis Falasco. United states patent classification: system organization. *World Patent Information*, 24:111 – 117, 2002.
- [Fel98] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.
- [Fog07] Pasquale Foglia. Patentability search strategies and the reformed ipc: A patent office perspective. *World Patent Information*, 29:33–53, 2007.
- [Gat] Developing language processing components with gate. <http://gate.ac.uk/sale/tao/>, last visited: 2009-12-08.
- [Hep00] Mark Hepple. Independence and commitment: assumptions for rapid training and execution of rule-based pos taggers. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Hong Kong, China*, pages 278–277. Association for Computational Linguistics, 2000.
- [HNR07] David Hunt, Long Nguyen, and Matthew Rodgers, editors. *Patent Searching: Tools & Techniques*. Wiley, 2007.
- [IAS07] Kishore Varma Indukuri, Anurag Anil Ambekar Ambekar, and Ashish Sureka. Similarity analysis of patent claims using natural

- language processing techniques. In *Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'07)*, Sivakasi, India, pages 169–175. IEEE Computer Society, 2007.
- [IPC09] International patent classification guide. http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc_2009.pdf, last visited: 2009-12-08, Version 2009.
- [KR09] Kas Kasravi and Marie Risov. Multivariate patent similarity detection. In *Proceedings of the 42nd Hawaii International Conference on System Sciences (HICSS)*, Waikoloa, Hawaii, pages 1–8. IEEE Computer Society, 2009.
- [Lar99] Leah S. Larkey. A patent search and classification system. In *Proceedings of the fourth ACM conference on Digital libraries (DL'99)*, Berkeley, California, pages 179–187. Association for Computing Machinery, 1999.
- [Mic06] Jacques Michel. Considerations, challenges and methodologies for implementing best practices in patent office and like patent information departments. *World Patent Information*, 28(2):132–135, 2006.
- [MIW05] Andreea Moldovan, Radu Bot Ioan, and Gert Wanka. Latent semantic indexing for patent documents. *International Journal of Applied Mathematics and Computer Science*, 15:551–560, 2005.
- [MMO⁺05] Hisao Mase, Tadataka Matsubayashi, Yuichi Ogawa, Makoto Iwayama, and Tadaaki Oshio. Proposal of two-stage patent retrieval method considering the claim structure. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4:190–206, 2005.
- [MPE08] Manual of Patent Examination Procedure (MPEP). <http://www.uspto.gov/web/offices/pac/mpep/mpep.htm>, last visited: 2009-12-08, July 2008.
- [Phi05] Minoo Philipp. Why pay for value-added information? *World Patent Information*, 27:7–11, 2005.
- [RM95] Lance Ramshaw and Mitchell Marcus. Text chunking using transformation-based learning. In *Proceedings of the Third*

- Workshop on Very Large Corpora (WVLC), Cambridge, Massachusetts*, pages 82–94. Association for Computational Linguistics, 1995.
- [RSG08] James F. Ryley, Jeff Saffer, and Andy Gibbs. Advanced document retrieval techniques for patent research. *World Patent Information*, 30:238–243, 2008.
- [Sch00] Paul Schwander. An evaluation of patent searching resources: comparing the professional and free on-line databases. *World Patent Information*, 22:147–165, 2000.
- [She03] Svetlana Sheremetyeva. Natural language analysis of patent claims. In *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing, Sapporo, Japan*, pages 66–73. Association for Computational Linguistics, 2003.
- [SHWA03] Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003), Sapporo, Japan*, pages 8–15. Association for Computational Linguistics, 2003.
- [Sim06] Edlyn S. Simmons. Patent databases and gresham’s law. *World Patent Information*, 28:291–293, 2006.
- [SOMI03] Akihiro Shinmori, Manabu Okumura, Yuzo Marukawa, and Makoto Iwayama. Patent claim processing for readability: Structure analysis and term explanation. In *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing, Sapporo, Japan*, pages 56–65. Association for Computational Linguistics, 2003.
- [TFI04] Toru Takaki, Atsushi Fujii, and Tetsuya Ishikawa. Associative document retrieval by query subtopic analysis and its application to invalidity patent search. In *Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management (CIKM), Washington D.C.*, pages 399–405. Association for Computing Machinery, 2004.
- [TLB⁺09] John Tait, Mihai Lupu, Helmut Berger, Giovanna Roda, Michael Dittenbach, Andreas Pesenhofer, Erik Graf, and Keith Van Rijsbergen. Patent search: An important new test bed for IR.

- In *Proceedings of the 9th Dutch-Belgian Information Retrieval Workshop (DIR 2009)*, Enschede, Netherlands, 2009.
- [WIP] Understanding industrial property. http://www.wipo.int/freepublications/en/intproperty/895/wipo_pub_895.pdf, last visited: 2009-12-08.
- [YS08] Shih-Yao Yang and Von-Wun Soo. Comparing the conceptual graphs extracted from patent claims. In *Proceedings of the 2008 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC 2008)*, Taichung, Taiwan, pages 394–399. IEEE Computer Society, 2008.