Chapter 10
Patent Claim Decomposition for Improved Information Extraction

Peter Parapatics and Michael Dittenbach

Abstract In several application domains research in natural language processing and information extraction has spawned valuable tools that support humans in structuring, aggregating and managing large amounts of information available as text. Patent claims, although subject to a number of rigid constraints and therefore forced into foreseeable structures, are written in a language even good parsing algorithms tend to fail miserably at. This is primarily caused by long and complex sentences that are a concatenation of a multitude of descriptive elements. We present an approach to split patent claims into several parts in order to improve parsing performance for further automatic processing.

10.1 Introduction

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 invention described more comprehensible. Both, the European[1] as well as the US definition[2] of patent claims put emphasis on conciseness and clarity. This and further official guidelines on claim formulation have several implications on the language used. In this work, we investigate how the structure of the claims-specific language can be used to split them into several components and rearrange them in order to improve the performance of natural language processing tools such as dependency parsers and to improve readability. To this end, we use the English language parts of a set of European patent documents from the International Patent

[1] http://www.epo.org/patents/law/legal-texts/html/epc/2000/e/ar84.html.

[2] http://www.gpoaccess.gov/uscode/browse.html.

P. Parapatics (✉)
Department of Software Technology and Interactive Systems, Vienna University of Technology, Favoritenstr. 9-11/188, 1040 Vienna, Austria
e-mail: p.parapatics@gmail.com

M. Dittenbach
max-recall information systems, Vienna, Austria
e-mail: m.dittenbach@max-recall.com

M. Lupu et al. (eds.), *Current Challenges in Patent Information Retrieval*,
The Information Retrieval Series 29,
DOI 10.1007/978-3-642-19231-9_10,

Classification (IPC) category A61C (Dentistry; Oral or Dental Hygiene). The goal of this research is the development of a method to automatically decompose the often long and winding sentences into smaller parts, identifying their constituents and relations and putting them into a machine-processable structure for further analysis and visualization.

10.2 Patent Claim Structure

In general, rules for examining, and thus also for drafting a patent are quite similar internationally, but there are variations from patent office to patent office. The characteristics described in this paper are based on the Guidelines for Examination in the European Patent Office (EPO) as of April 2009, Part C, Chap. III [2] and the Manual of Patent Examining Procedure of the United States Patent and Trademark Office (USPTO) [4]. The EPO as well as the USPTO require every patent document to contain one or more claims. The claims section is the only part of a patent conferring protection to the patent holder. The description and drawings should help the examiner to understand and interpret the claim but do not provide any protection themselves. Due to the importance of the claims there are very precise syntactic and semantic rules that have to be followed when drafting patent claims. A patent contains one or more independent claims that define the scope of the invention [2, Sect. 3.4]. Additionally, a patent may contain dependent claims which impose further limitations and restrictions on other dependent or independent claims. Each claim has to be written in a single sentence.

Independent claims should start with a part which describes already existing prior art knowledge and is used to indicate the general technical class of the invention. It describes the elements or steps of the invention that are conventional or known. These are then refined in a part describing the aspects or steps of the invention which are considered new or improved and which the patent holder wants to protect. These two parts are connected with specific key phrases which vary between the USPTO and EPO. Moreover, the terminology for naming the parts differs slightly. The USPTO refers to the part describing prior art as *preamble* [4, Chap. 608.01(i)]. The key phrase is called *transitional phrase* and the main part of the claim is referred to as the *claim body*. In the transitional phrase, keywords such as “comprises”, “including” or “composed of” are used. The EPO suggests the same claim structure but does not name the separate parts [2, Sect. 2.2]. It refers to this structure as the *two-part form* (not counting the transitional phrase) with the first part corresponding to the preamble and the second part to the claim body. The two parts are linked with either the phrase “characterized by” or “characterized in that”.

Independent claims do not necessarily have to be defined in the two-part form. The EPO [2, Sect. 2.3] considers the two-part form inappropriate for claims which describe:

- 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;
- or a complex system of functionally inter-related parts, the inventive step concerning changes in several of these or in their inter-relationships.

An example claim for the third rule is the following claim taken from a patent document in the dentistry domain: "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."

A dependent claim can refer to independent as well as other dependent claims and are used to refine and describe additional details or parts of the invention. It has to incorporate all features from the claim it refers to and must not broaden the previous claim. 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 the refinement or the definition of parts of the invention. The two-part form, where the two parts are linked with "characterized in that" or "characterized by", is not required for dependent claims but is nevertheless very common. Other common link phrases between the two parts are "wherein" and "comprising" such as in the claims "The orthodontic bracket of claim 1 *wherein* said bracket is [...]" or "An apparatus usable for carrying out the method according to claim 1 or 2, *comprising* [...]"

The USPTO explicitly defines rules for the order of claims in the patent [4, Chap. 608.01(n)]. In the EPO guidelines the order is stated implicitly. Dependent claims have to be ordered from the least restrictive to the most restrictive. This is important from a machine processing point of view, in the sense that concepts or terms which are refined in a dependent claim have already been introduced in a preceding claim in the document.

Claims have a different form depending the type of invention they describe. It can be differentiated between claims to physical entities (product, apparatus) and claims to activities (process, use) [2, Sect. 3]. *Product* and *apparatus claims* normally have the following form: "An X, comprising a Y and a Z". *Method claims* have a very similar form but instead of describing parts of a physical entity a sequence of steps are described. "A method for X comprising (the steps of) heating Y and cooling Z". A *use claim* is usually written in the following form: "The use of X for the Z of Y".

Several common grammatical structures can be found in patent claims. One that is commonly used in claims is an enumeration of several parts of prior art improvements or steps of a method. These enumerations occur in various syntactic forms like: "An M comprising an X, a Y and a Z" or "An M comprising: (a) an X, (b) a Y and (c) a Z".

Since a claim should be as concise as possible (cf. [2, Sect. 4]), each term used in the claim must have a definite and unambiguous meaning. New concepts are introduced with an indefinite article ("a" or "an"). Subsequent uses of the same element are preceded by "the" or by "said".

10.3 Related Work

Research is done in various fields of patent processing.

In [7] the authors aim to quantify three challenges in patent claim parsing: claim length, claim vocabulary and claim structure. Their experiments show that the average sentence length of claims is longer compared to general English sentences even if the claims are split on semicolons and not only on full stops. This results in more structural ambiguities in parses of long noun phrases. While the vocabulary is similar to normal English texts the authors show that the distribution of words does differ. The biggest challenge for syntactic parsing poses the sentence structure as claims consist of sequences of noun phrases rather than clauses.

The authors of [3] propose a technique for claim similarity analysis which could be used for building patent processing tools to support patent analysts. They compute a similarity score between two claims based on simple lexical matching and knowledge based semantic matching. The syntactic similarity measure is based on the number of nouns that occur in both claims. For semantic similarity a score is computed by comparing each noun from the first claim to all nouns from the second claim using WordNet [1]. The highest score is recorded. The final semantic similarity score for two claims is then calculated by summing up the semantic similarity score for each noun.

A complex and domain-specific NLP-based approach is used in [5]. It is claimed that the use of broad coverage statistical parsers like the Stanford Natural Language Parser[3] is not appropriate for the patent domain. Since they are trained on general language documents, the accuracy of these parsers suffers when used for parsing patent claims. The proposed parsing method relies on supertagging and uses a domain-specific shallow lexicon for annotating each lexeme with morphological, syntactic and semantic information. Semantic information consists of an ontological concept defining the word membership in a certain semantic class (Object, Process, etc.). In the supertagging procedure each word is annotated with several matching supertags. In the following disambiguation procedure, hand crafted rules are used to eliminate incorrect supertags. The central part of the method is the predicate lexicon which is used to create a predicate-argument structure by annotating each predicate with syntactic and semantic information. A grammar is used to fill each argument of a predicate with a matching chunked phrase (e.g.: NP, NP and NP) from the claim based on the syntactic and semantic information in the supertag.

In [8] patent claims are compared by computing a similarity measure for conceptual graphs extracted from the claims using a natural language parser. A conceptual graph G is a set of (C, R, U, lab) where C are the concept vertices, R the relation vertices, U a set of edges for each relation. A label from the set lab is assigned to every vertex in the graph. A specific domain ontology is used for the concept and relation vertices in the conceptual graph. The conceptual graphs are extracted from dependency relations created with the Stanford Parser. The developed method is intended to be used for infringement searches and in particular for tasks such as patent clustering, patent comparison and patent summarization.

[3] http://nlp.stanford.edu/software/lex-parser.shtml.

Table 10.1 Data sets: characteristics

Data Set	Claim type	Nr. claims	Nr. words	Avg. claim length
Analyzed Set	Ind. claims	159	20, 321	127.81
	Dep. claims	862	28,794	33.40
Evaluation Set	Ind. claims	13,628	1,803,341	132.33
	Dep. claims	73,706	2,415,533	32.77

The authors of [6] focus on structural analysis of Japanese patent claims in order to create parsing methods for specific claim characteristics. They show that Japanese patent claims are very similar to European and US claims in the sense that a single sentence out of multiple sentences using specific keywords and relations. Six common relations (Procedure, Component, Elaboration, Feature, Precondition, Composition) are described which can be found in Japanese patents. These relations can be identified by cue phrases, for which a lexical analyzer is used in order to decompose a patent claim into several parts.

10.4 Data Set

For creating and evaluating our method, which will be described in the next section, 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 set.[4] Table 10.1 shows the characteristics of the two data sets. The figures show that independent claims are more than three times as long as dependent claims.

Table 10.2 shows the success rate (coverage) of the Stanford parser applied to the claims. 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. The coverage provides a good indication for the complexity of a text. The higher complexity of independent claims is therefore underlined by the high number of unsuccessful parses of independent claims as compared to dependent claims. 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 amount of memory (JVM max. heap size). This is an important parameter, because of the memory requirements for constructing the large parse trees for the relatively long independent claims. An infor-

[4] http://ir-facility.org.

Table 10.2 Stanford parser success rate

Data Set	Claim type	JVM max. heap size	Successful parses	Failed parses	% of successful parses
Analyzed Set	Ind. claims	1000 MB	132	27	83.01%
		500 MB	89	70	55.97%
	Dep. claims	1000 MB	859	3	99.65%
		500 MB	848	14	98.38%
Evaluation Set	Ind. claims	1000 MB	10,671	2,957	78.30%
		500 MB	7,482	6,146	54.90%
	Dep. claims	1000 MB	73,427	279	99.62%
		500 MB	72,769	937	98.73%

mal evaluation of the parse trees indicates that the quality of the results is very low for the long and complex claim sentences.

10.5 Method

10.5.1 Preprocessing

Before a patent document is decomposed, a number of data preprocessing and cleaning steps are executed to normalize the claim text. 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)". For our purpose, 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 (but retaining mathematical and chemical formulas):

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

In some claims, elements of an invention are enumerated in a form such as "a." or "b.". 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 and are therefore removed.

In many documents the actual claim text is preceded by its claim number. Since this information is already implicitly given via the order of the claims in the patent document it is removed.

The term "characterized" is an important element that needs to be identified. The British spelling variant is replaced by the American one.

In the last preprocessing step all occurrences of the word "said" are replaced with the definite article "the". This is a simple but effective way of improving the

Table 10.3 Claim Types

Data Set	Claim type	Number of claims
Analyzed Set	Physical Entity Claims	114
	Method Claims	41
	Use Claims	4
Evaluation Set	Physical Entity Claims	10,310
	Method Claims	3,315
	Use Claims	3

performance of natural language parsers even before decomposing the claims. Natural language parsers trained on general language texts interpret the word "said" as a verb. In claims, however, it is always used for referring to an already introduced concept.

10.5.2 Claim Type and Category Identification

A simple heuristic is used to determine whether a claim is dependent or independent. The drafting guidelines for dependent claims suggest that it 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" are classified as dependent claims, all others as independent claims.

Independent claims can be categorized into: *physical entity claims*, *method claims*, *use claims*. This distinction is important, because the types differ slightly and require distinct analysis patterns. A 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 patent domain.

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 and all claims which start with the phrase "The use" are classified as use claims. Thus, simple string matching can be used.

No such simple heuristics are available for identifying physical entity claims. Physical entity claims usually start with the claimed invention rather than with claim-specific keywords. Claims that can neither be classified as use claims nor as method claims are classified as physical entity claims.

Table 10.3 shows the frequency of each claim category in the two data sets. The figures show that the number of physical entity claims is about three times higher than the number of method claims and it can also be seen that almost no use claims are present in the data sets.

10.5.3 Claim Decomposition

Our process of decomposing claims consists of three main phases: pattern identification, pattern extraction, post processing and merging the extracted parts into a tree structure. 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. Therefore, the claims are analyzed with GATE[5] an open source natural language processing framework. Each claim is tokenized and a sentence-splitter is applied. Depending on the requirements of the extraction rules, Parts-Of-Speech tagging and Noun Phrase Chunking is done.

Based on the annotations created by the rules (JAPE grammars) the claims can be decomposed. For this purpose the textual 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 to remove unnecessary characters such as white spaces, punctuation symbols and words from the extracted parts.

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 relation type to the parent. Each node contains the text of the extracted part and, 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.

10.5.4 Independent Claim Decomposition

Due to space considerations, we focus more on the decomposition of independent claims in this article, since they are longer and more complex than dependent claims and thus more interesting. 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. In the following section the extraction rules for physical entity claim are described.

10.5.4.1 General Patterns

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

[5] http://gate.ac.uk/.

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 can 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 1. Therefore each claim sentence is analyzed with GATE and the first noun phrase is extracted as claim-subject.

Example 1 (EP1444966-A1)

$$\overbrace{\text{A dental head unit}}^{\text{Claim-Subject}} \text{ capable of measuring a root canal length of a patient}$$

Characterized-Pattern If a claim is drafted in the two-part form as suggested by the EPO, the keyphrases "characterized in that" and "characterized by" can be used to split the claim into 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 Sect. 10.5.4.2. 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.

10.5.4.2 Physical Entity Claims

The focus in this method was set on the analysis of physical entity claims. Due to the comparatively large number of physical entity claims in the Analyzed Set, it was possible to identify a larger number of patterns.

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 the introduction of a new concept. The parts can be identified by looking for singular or plural noun phrases preceded by the indefinite article "a" or "an" such as shown in Example 2.

Example 2 (EP0063891-B2)

Claim-Subject

An ejector holder [...] , the holder comprising

Composition-Start

Composition-Part Composition-Part

an elongate barrel [...], a plunger [...], [...] and a lever [...].

Composition-Part

The JAPE grammar used for extracting Composition-Patterns first annotates the start of a Composition-Pattern by looking for the keywords mentioned above. It then identifies and annotates the composition-parts the Composition-Pattern is composed of by looking for singular or plural noun phrases preceded by the indefinite article "a" or "an". The grammar takes into account several different linguistic patterns in order to identify nested Composition-Patterns. Each extracted part is attached to the CLAIM-SUBJECT 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 to create one single sentence. A very common structure used for this purpose is the Nested-Sentence-Pattern where an already introduced concept is refined. Example 3 shows the typical structure of a Nested-Sentence-Pattern.

Example 3 (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. A nested sentence ends when either another Nested-Sentence-Pattern is found or the sentence ends. 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 found in the claim (Nested-Sentence or Composition-Pattern), are extracted as description-part. The description usually indicates the purpose of the invention (see Example 4). In some cases, however, it describes elements that an invention contains.

Example 4 (EP0415508-A2)

Claim-Subject

An apparatus to continuously harden light curing resins, comprising [...]

Description

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

10.5.4.3 Characterized-Part Decomposition

If the claim is drafted in the two-part form as 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. To this end, the Composition-Pattern rules are used in a slightly modified version. 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 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.

10.5.4.4 Composition-Part Decomposition

Extracted composition-parts can be further decomposed 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 5.

Example 5 (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 like verbs in gerund form possibly preceded by an adverb ("a neck section *extending proximally* from the head section [...]") or verbs in past tense, possibly preceded by an adverb ("a brush part *detachably attached* to one end of the drive shaft"). 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 using the Nested-Sentence-Pattern rule.

10.5.5 Dependent Claim Analysis and Decomposition

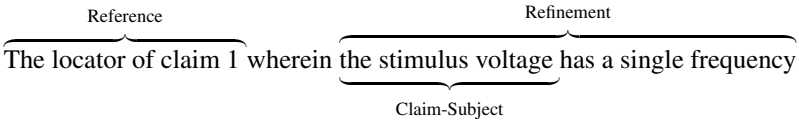
Dependent claims consist of two parts. The first part provides a reference to the claim(s) it refines while the second part describes the refinement itself. The analysis of dependent claims consists of two tasks. In the first step the reference-part is analyzed to extract the references to refined claims. References to previous claims 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. The most important ones are single numbers and ranges of numbers preceded by the word claim such as in "The locator of *claim 1* wherein [...]" or "An article as claimed in any of *claims 12 to 14*, wherein [...]". The annotated references 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. These references can than 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 those used for decomposing independent claims.

First the claim is split into two parts, the reference-part and the refinement-part. A JAPE grammar is used to identify the end of the reference-part according to several linguistic patterns. In the most commonly used pattern the reference-part ends with one of the phrases ", wherein", ", characterized in that" or "characterized by" such as in "Hinge member as claimed in claim 1*, wherein* the head means is circular [...]".

Then, 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 6 provides a better understanding of the claim-subject extraction rule. 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.

Example 6 (EP0171002-B1)



Finally 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 rules used for decomposing characterized-parts from physical entity claims (see Sect. 10.5.4.2).

10.5.6 Merging of Dependent and Independent Claims

After the claims have been analyzed and decomposed, a coreference resolution 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. 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, the fact that a new element is usually introduced with a phrase such as "a CONCEPT" and later referred to as "the CONCEPT" can be exploited. For each claim a concept index containing *New-Concepts* and *Ref-Concepts* is created. The merging algorithm is illustrated in Example 7, showing the decomposition of an independent and a dependent claim and 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", introducing the refined element in the independent claim.

Example 7 (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 [...]
```

Merged claims:

```
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 [...]
```

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.

10.6 Evaluation

10.6.1 Independent Claim Decomposition

In this section it is evaluated how the method developed in this work reduces the length and complexity of independent claims. To this end the average length of the original independent claims is compared with the average length of parts extracted from these claims. The coverage of the Stanford Parser is used as a measure for complexity reduction. In order to provide an estimation of the quality of the rule

Table 10.4 Length reduction: independent claims

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 10.5 Length reduction comparison for claim categories

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

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 by manually classifying the claims as "correct/mostly correct", "partly correct" and "incorrect/insufficiently decomposed".

Table 10.4 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 cannot only be applied to European patents but can also handle the structurally slightly different US patents.

Table 10.5 compares 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 for both data sets. 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.

The achieved complexity reduction can be estimated from the number of successful parses using the Stanford Parser. Table 10.6 shows the success rate of the

Table 10.6 Stanford parser success rate: extracted parts

Data set	JVM max. heap size	Successful parses	Failed parses	% of successful parses	Improvement
Analyzed Set	1000 MB	1,010	2	99.80%	**16.79%**
	500 MB	1,003	9	99.11%	**43.14%**
Evaluation Set	1000 MB	100,140	151	99.85%	**21.55%**
	500 MB	99,793	498	99.50%	**44.60%**

Table 10.7 Quality estimation: physical entity claims

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

Table 10.8 Quality estimation: method claims

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

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. The last column shows the improvement compared to applying the parser to the original claims. The comparison shows that the coverage of the Stanford Parser is significantly higher on the extracted parts than on the original claims with the improvement being even slightly higher on the Evaluation Set.

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 10.7. From the 15 analyzed claims nine are decomposed correctly or almost correctly, two are considered partially correct and four are classified as incorrect or insufficiently decomposed.

Table 10.8 shows the evaluation results for the 10 analyzed method claims. The figures show that four claims are decomposed correctly, two are partially correct and four 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

Table 10.9 Resolved claim references

		Total Number	Percentage
Analyzed Set	Attached claim references	81	96.43%
	Missing claim references	3	3.57%
	Total number of dependent claims	84	100%
Evaluation Set	Attached claim references	77	100%
	Missing claim references	0	0%
	Total number of dependent claims	77	100%

for carrying out the method or enumerate steps in a form that cannot be handled correctly by the rules.

10.6.2 Claim Merging

From each of the data sets, 10 patents containing a physical entity claim were randomly selected 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 it was not identified by the decomposition rules.

Table 10.9 shows 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 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 successfully extracted for two dependent claims.

Table 10.10 provides 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.

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

Table 10.10 Attachments

		Total Number	Percentage
Analyzed Set	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%
Evaluation Set	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%

by the rules. The figures show that the number of correct attachments is relatively high while there are almost no incorrect attachments. The figures also show that the percentage of parts for which no attachment was found is relatively high in both data sets. One reason is that a *Ref-Concept* in a dependent claim can be provided in a shorter form than the original *New-Concept* as for example a concept may be introduced as "spaced-apart arms" in an independent claim and referenced with "the arms" in the dependent claim.

Another reason is that some dependent-claim-subjects are not extracted correctly due to erroneous POS-tagging. This affects especially the term "means". This occurs for phrases such as "The impression tray according to claim 1 in which the *light-reflecting* means comprises a thin layer of reflective metal.". In this case 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 is followed for extracting composition-parts.

The third reason is that the extracted claim-subject is not always the concept which is refined. This is shown in the phrase "The impression tray according to claim 5 in which *the edges* of *the cover sheet* are sealed to [...]" where the term "the edges" is extracted as claim-subject instead of the words "the cover sheet".

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 as in the claim "A teeth straightening bracket according to claim 1 characterized in that engaging fingers [...] are disposed except for the both longitudinal ends of *the wire support*".

10.7 Conclusions and Future Work

We have shown that the automatic analysis of patent claims using natural language parsers can be dramatically improved by decomposing them first into smaller units using a set of rules and heuristics. This research is a first step toward developing sophisticated methods and tools to facilitate the work of patent information professionals by automatically analyzing, structuring and visualizing patent claims.

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 needs to 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 gold standards and precise criteria for measuring the quality and the correctness of the extracted claim parts.

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.

References

1. Fellbaum C (ed) (1998) WordNet: An electronic lexical database (language, speech, and communication). MIT Press, Cambridge
2. 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
3. Indukuri KV, Ambekar AAA, Sureka A (2007) 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. IEEE Computer Society, Washington, pp 169–175
4. Manual of Patent Examination Procedure (MPEP) (2008) http://www.uspto.gov/web/offices/pac/mpep/mpep.htm, last visited: 2009-12-08
5. Sheremetyeva S (2003) Natural language analysis of patent claims. In: Proceedings of the ACL-2003 workshop on patent corpus processing, Sapporo, Japan. Association for Computational Linguistics, Stroudsburg, pp 66–73
6. Shinmori A, Okumura M, Marukawa Y, Iwayama M (2003) Patent claim processing for readability: Structure analysis and term explanation. In: Proceedings of the ACL-2003 workshop on patent corpus processing, Sapporo, Japan. Association for Computational Linguistics, Stroudsburg, pp 56–65
7. Verberne S, D'hondt E, Oostdijk N, Koster CH (2010) Quantifying the challenges in parsing patent claims. In: Proceedings of the 1st international workshop on advances in patent information retrieval (AsPIRe 2010), Milton Keynes, UK, pp 14–21
8. Yang S-Y, Soo V-W (2008) 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. IEEE Computer Society, Washington, pp 394–399