A Manually Annotated Corpus of Pharmaceutical Patents*

Márton Kiss[1], Ágoston Nagy[1], Veronika Vincze[2,3],
Attila Almási[1], Zoltán Alexin[4], and János Csirik[2]

[1] University of Szeged, Department of Informatics
6720 Szeged, Árpád tér 2., Hungary
[2] MTA-SZTE Research Group on Artificial Intelligence
6720 Szeged, Tisza Lajos krt. 103., Hungary
[3] Universität Trier, Linguistische Datenverarbeitung
54286 Trier, Universitätsring, Germany
[4] University of Szeged, Department of Software Engineering
6720 Szeged, Árpád tér 2., Hungary
{mkiss,nagyagoston,vinczev,alexin,csirik}@inf.u-szeged.hu,
vizipal@gmail.com

Abstract. The language of patent claims differs from ordinary language to a great extent, which results in the fact that tools especially adapted to patent language are needed in patent processing. In order to evaluate these tools, manually annotated patent corpora are necessary. Thus, we constructed a corpus of English language pharmaceutical patents belonging to the class A61K, on which several layers of manual annotation (such as named entities, keys, NucleusNPs, quantitative expressions, heads and complements, perdurants) were carried out and on which tools for patent processing can be evaluated.

Keywords: patent, corpus, syntactic annotation, named entities.

1 Introduction

For the automatic processing of patents, they are required to be linguistically preprocessed, that is, to be tokenized, POS-tagged and syntactically parsed. However, the language of patent claims differs from ordinary language to a great extent, which results in the fact that tools especially adapted to patent language are needed in patent processing, what is more, manually annotated corpora are desirable to evaluate the performance of these tools.

Thus, we constructed a toolkit that is able to split English language patents into sentences (clauses), to parse them morphologically and to identify key concepts such as named entities or keywords in the texts. In order to evaluate the performance of our tools, we constructed a corpus of pharmaceutical patents belonging to the class A61K, on which several layers of manual annotation were carried out. In this paper, we present our corpus and offer a detailed description of the manual annotations.

* This work was supported in part by the National Innovation Office of the Hungarian government within the framework of the project MASZEKER.

P. Sojka et al. (Eds.): TSD 2012, LNCS 7499, pp. 135–142, 2012.

2 Motivation and Related Work

The claims section of patents contains usually the most important information about the topic and the scope of the patents. Among the claims it is the main claim that summarizes the essential content of the patent: all the necessary characteristics of the method, process, tool or product described in the patent have to be listed here. The other claims further detail these characteristics, often with the help of figures, tables and images [1].

2.1 The Linguistic Characteristics of Patent Claims

The linguistic features of patent claims considerably differ from those of ordinary language. The main claim typically consists of one very long sentence with a complex syntactic structure, which is quantitatively supported by the experiments described in [2] and in [3]. There are multiple embedded clauses and noun phrases, lists and coordinated phrases in main claims. Elliptic constructions, anaphoras, post-head modifiers and relative clauses also make the automatic processing of patent claims difficult.

The vocabulary of patents also contains neologisms: it is mostly multiword expressions (noun compounds) that cannot be found in general dictionaries [3]. However, they are compositional, i.e. their meaning can be calculated from the meaning of their parts and from the way they are connected. Sometimes it is also a source of problem that many words acquire a new meaning within the patent since the process or product described is also a novelty, hence it may well be the case that old terms are used in a slightly modified meaning [1].

As authors are required to provide a very detailed description of the subject of the patent, the language used is strict and precise. Still, there is a tendency to generalize over the scope of the patent in order to prevent further abuse [1]. Thus, the scope of the patents can be expanded or other use cases can later be included in the patent. The linguistic strategies applied include the following:

- the use of *etc.* at the end of lists or enumerations;
- the use of *for instance* or *e.g.* at the beginning of lists or enumerations;
- the use of inclusive *or*;
- the use of generalizing adverbs (*usually*, *typically* etc.)

These strategies are comparable to uncertainty cues, in other words, hedges or weasels [4]. Nevertheless, whereas the use of hedges and weasels and other vague and misleading phrases is undesirable in e.g. Wikipedia articles, their frequent occurrence in patent claims is a general phenomenon.

2.2 Related Work

There have been several patent corpora constructed. For instance, the European Patent Corpus contains 130 million sentence pairs from 6 languages, in which sentences are automatically aligned [5]. There is a Japanese–English patent parallel corpus containing 2 million pairs of aligned sentences [6] and a Chinese–English patent parallel corpus has also been constructed with 14 million sentence pairs [7] to name but a few. These corpora can be effectively exploited in cross-lingual information retrieval and machine translation tasks.

The linguistic characteristics of patent claims call for special techniques to be applied when processing the claims. In order to evaluate the tools adapted to patent processing, manually annotated data are needed. However, with the exception of the 100 sentences annotated by [3], we are not aware of any manually annotated patent corpora, which motivated us to build a corpus with several layers of manual annotation on which our processing toolkit can also be evaluated. In the following sections, our corpus and toolkit are presented.

3 The Corpora

We collected 10,000 patents (C10K) (see below) out of which we randomly selected 313 patents (C313) from the class A61K, which includes preparations for medical, dental or toilet purposes and we later narrowed them down to 62 claims (C62). The latter corpus has been chosen to be our benchmark database as it contains all types of manual annotations to be described in this section. Table 1 shows the main characteristics of the main claims of the corpora.

Table 1. Comparison of the corpora

	C62	**C313**	**C10K**		**C62**	**C313**	**C10K**
Patent	62	313	8,797	Text	●	●	●
Sentence	62	865	8,793	Key	●		
Token	7,883	59,356	1,771,290	NucleusNP	●	●	
Lemma	1,466	6,010	32,252	Quantity	●	●	
NucleusNP	1,706	14,275		Dependent	●		
Perdurant	664	3,448		Perdurant	●	●	
Quantity	226			Enumeration	●		
Key	415			Headword	●		
NE	825	20 374		Named entity	●	●	

The 10K Corpus. For a start we prepared a corpus of 10,000 patents. The corpus is made of patents chosen randomly from 10 different IPC subclasses (A24F, A61K, A63K, B26D, D21F, E01D, F21K, G10C, G10L, H04M). We downloaded 1,000 patents for each of the above mentioned subclasses from the website of the United States Patent and Trademark Office[1]. We think that this hierarchy level is appropriate for our research, thus we did not carry out further segmentations within the subclasses. Since each patent is downloadable in a well-formatted full-text format from the site of the United States

[1] http://patft.uspto.gov/

Patent and Trademark Office, we could easily retrieve the required information, which we converted to XML format. The corpus in XML format was easily manageable in the UIMA[2] system.

The C313 Corpus. The C10K proved too big for the task of annotation, so we filtered it. We chose 313 patents that belong to the subclass A61K. We did initial research on this smaller corpus. On the 313 A61K patents, the following annotations were manually marked: named entities (NEs), NucleusNPs, perdurants, quantitative expressions. When generating the verb frames, the verb frames of the verbs of this C313 corpus were considered.

The C62 Corpus. For marking specific annotations, the C313 corpus seemed enormously big as well. So we constructed a corpus of 62 patents from the 313. We carried out the markings of the enumerations and the keys only on C62.

4 Manual Annotation of the Corpora

In this section, we describe the linguistic phenomena that are manually annotated in our corpora.

4.1 Keys

The main claim of a patent is usually a very complex sentence including many subordinations and coordinations, which is difficult to analyze. To analyze these sentences by the current automated algorithms is hardly possible [3] hence we need to find a solution to break these sentences into sentence fragments that are analyzable by automated algorithms. Therefore we marked the postmodifiers and the beginnings of clauses with keys.

Keys are generally the sections of the processed text where the presence of the modifier-modified noun relation is purely recognizable on formal grounds. Keys consist of a first and a second part. Shinmori et al. [2] apply a similar technique to break Japanese patents into analyzable fragments, however, they only mark the beginning of clauses (i.e. the second part of our keys). **Simple keys** serve to indicate successive keys if no remote second-type key is connected to the first part of the key. E.g.: *substance which*, *group consisting*. The key is **complex** if the first and the second part of the key are not directly following each other or if more second parts belong to the first part of the key. E.g.: *the **process comprising** the steps of deforming the films (18) to form a multiplicity of recesses (16), **filling** the recesses.*

Keys were marked in C62 by hand to help the development of the automatic key marker module.

[2] UIMA means *Unstructured Information Management Architecture*, and is used in this project to give structure to unstructured documents (e.g. plain texts) by different, user-defined or already existing external modules performing annotation tasks and to visualize these annotations in a user-friendly way. `http://uima.apache.org/`

4.2 NucleusNPs and Their Nominal Heads

In the corpus, it is not the standard NP projection – well-known from generative syntax, e.g. [8] – that is used but another one that we named **NucleusNP**. The main difference between the two types of projection is that a standard NP can have complements and postnominal adjectival adjuncts attached to its head, which is not allowed for NucleusNPs. The nominal head of NucleusNPs marks their end boundary (if it is not followed by a quantitative expression), thus eventual prepositional phrases are not attached to it. Therefore, NPs not having a prepositional complement or a postnominal adjectival phrase coincide with NucleusNPs (e.g. *ascorbic acid*), but the others do not. To sum up, a NucleusNP obligatorily has a nominal head, and optionally prenominal adjectival phrases as well as pre- or postnominal quantitative expressions.

The manual annotation coincides with the above mentioned definition of NucleusNPs. In the following example all NucleusNPs are marked (in italics), with their nominal head (underlined) and quantitative expression (bold).

> *A pharmaceutical composition* comprising [...] in *a ratio* of *paracetamol* to *calcium carbonate* of **3.0:1.0 to 30.0:1.0**, ***at least one*** *binding agent*, [...] ***at least 60%*** *of the paracetamol* is released from *the composition* at **180 seconds** [...] at **40°C.±2°C.** [...]

NucleusNPs are manually marked in C62 and C313.

4.3 Quantitative Expressions

As the corpus describes many chemical compositions, and the ingredients of these have to be detailed as precisely as possible, it possesses many quantitative expressions. As quantitative expressions are important for semantic document indexing, these have to be identified. For that purpose, quantitative expressions were manually annotated in the corpus; however, in the manual annotation phase, these quantitative phrases are marked as a whole, their internal structure is not annotated. The quoted example in the previous subsection showed some different quantitative expressions (marked in bold) annotated in the text.

As it can be seen from the example, quantitative expressions can be (1) intervals (*3.0:1.0 to 30.0:1.0*, *40°C.±2°C.*), (2) numbers with measure units (*180 seconds*), (3) numbers written with letters and eventual measure units (*at least one*), (4) relative expressions (*at least 60%*).

Quantitative expressions are manually annotated in C62 and C313.

4.4 Heads and Complements

The corpus was also annotated in terms of heads and complements, as well. Heads can be (1) finite or non-finite verb forms, (2) adjectival phrases and (3) nominal expressions. (1-3) may have complements introduced by a preposition (e.g. [$_{HEAD}$ *consisting*] [$_{OF-compl}$ *of ascorbic acid*]), only (1) can have complements not introduced by a preposition, which can precede or follow the head.

Heads were marked with bold letters, complements with italics, and of course, character series representing both are in bold and in italics in the same time. Complements are also labelled: this label can be found after the complement introduced by the _ sign. A default complement is attached to the nearest preceding head, and it is an *and*-type coordination; if it is not the preceding head, the number of the heads jumped out is marked between curly brackets { }. Here we show the final result of the annotation on a short text extract:

> **A blood sugar regulating product** ***obtainable***_mod1 *from soybean seeds*_from *by* ***a process***_by ***comprising***_mod1 [...] a) ***soaking***_obj *the soybean seeds*_obj [...] b) ***drying***{2} *the* [...] *seeds*_obj *to* **reduce**_to *the water content*_obj *of seeds*_of, c) ***grinding***{4} *the seeds*_obj [...]

This annotation scheme clearly shows the syntactic relation between heads and complements. For example, *by a process* is the prepositional complement of *obtainable*, and *a process* is the subject of the following verb *comprising*. *drying* and *grinding* are parts of the enumeration starting with *soaking*: these all are the direct objects of *comprising*: therefore, the first element is connected to *comprising*, and the other two to *soaking* as *and*-type coordinations – the numbers in curly brackets show the relative backward position of *soaking* as a head with respect to the actual complement (*soaking* is the second head counted backwards with respect to *drying*, and the fourth to *grinding*).

Heads and their complements are manually annotated in C62.

4.5 Perdurants

According to [9], perdurants – in contrast with endurants – are expressions that designate an event or a state, that is, they can fulfill the same functions as most verbs. So perdurant expressions can manifest themselves by a finite or non-finite verbal form (e.g. *prevents*, *preventing*), a deverbal adjective (e.g. *obtainable*), a noun derived from a verb (e.g. *to access* → *access*).

However, in our analysis, perdurant expressions are defined in another way: they are elements that can have prepositional adjuncts. Adjectives and the other nominal elements can only have prepositional elements that are in their respective valence pattern because it is verbs or perdurant expressions that are more likely to have prepositional adjuncts. Therefore, annotating perdurants facilitates parsing, where heads are connected to their complements because an unattached prepositional phrase is more likely to be linked to a perdurant expression or a verb than to a non-perdurant noun or adjective.

The following annotated text part shows all perdurant words (in italics):

> A tablet that readily *disintegrates* in gastric fluid to *give* aspirin crystals *coated* with a polymeric film [...] not *preventing access* of gastric fluid to the aspirin [...]

Perdurants are manually marked in the C62 and C313 corpora.

4.6 Other Annotations

Enumeration. In C62 we marked the enumerations by hand. We not only marked the type of the enumerations (discourse or linear) but also the borders of its items.

HeadWord. Pragmatically the subjects of the clause “we claim” (found at the beginning of the main claim) are the headwords. But in most cases this is a word of a very general meaning (*means*, *composition*, *method* etc.). A main claim can contain more than one headword. Headwords were marked on C62 as well.

Named entities. On C313 we marked the named entities. Since we examined the A61K subclass, we used the following labels when hand marking: disease (*drunkenness*), special (*succinic acid*), generic (*sugar*).

5 The UIMA Toolkit

During our research we used the UIMA linguistic framework. Now we show how we converted the documents containing manual annotations into the UIMA framework, as well as how we compared manual and automatic results. After that we describe the module that visualized results.

Manual Annotation and Word → UIMA converter. In order to facilitate linguistic annotation we have developed a converter which allows linguists to annotate using certain formatting in Word (e.g.: to change the background color of the marked text). Then we prepared UIMA annotations from the Word text to check and test machine algorithms.

Comparative Module for Annotations. In order to easily compare UIMA annotations with the manual annotations, we prepared a comparative module. The output of the comparison is the recall/precision/F-measure triplet, but we also generated lists for each comparison that contain which annotations were common and which were only included in either one or the other class of comparison.

Visualization. The UIMA annotations can be viewed with a visualizer created by us. This tool is able to visualize annotations and dependency trees as well. This module meant a big relief for the testing and troubleshooting section.

Parsing. After tokenizing and sentence splitting, the program performs the identification of chemical named entities, perdurants, quantitative expressions and NucleusNPs in the UIMA framework. After identifying these units we determine and connect heads with complements.

6 Conclusion

In this paper, we presented our corpora based on pharmaceutical patents belonging to the class A61K developed for the automatic linguistic processing of patents. The corpora contain several layers of manual annotation: keys, quantitative expressions, NucleusNPs, heads and complements, perdurants and named entities.

The development of the corpora contributes to the linguistic processing of patents, which makes it possible to develop applications in the field of information extraction / retrieval. For instance, the potential users of a search engine are usually interested in finding patents related to certain substances, treatments, illnesses etc. Thus, the identification of named entities is essential while further steps like the detection of perdurants are also necessary to extract events from the texts. Finally, for every application, the basic processing steps such as morphological analysis and syntactic parsing should also be carried out.

We would like to further develop our algorithms to recognize the linguistic phenomena annotated in the corpus in the future: its current modules may be ameliorated on the one hand and new modules may be implemented like an enumeration module on the other hand. We are also planning to adapt our toolkit for the processing of other patent classes such as electricity or transporting devices, therefore patent texts from these domains should also be annotated with the same layers.

References

1. Osenga, K.: Linguistics and Patent Claim Construction. Rutgers Law Journal 38, 61–108 (2006)
2. Shinmori, A., Okumura, M., Marukawa, Y., Iwayama, M.: Patent Claim Processing for Readability – Structure Analysis and Term Explanation. In: Proceedings of the ACL Workshop on Patent Corpus Processing, pp. 56–65. Association for Computational Linguistics, Sapporo (2003)
3. Verberne, S., D'hondt, E., Oostdijk, N., Koster, C.H.: Quantifying the Challenges in Parsing Patent Claims. In: Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval (AsPIRe 2010), pp. 14–21 (2010)
4. Farkas, R., Vincze, V., Móra, Gy., Csirik, J., Szarvas, Gy.: The CoNLL 2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL 2010): Shared Task, pp. 1–12. Association for Computational Linguistics (2010)
5. Täger, W.: The Sentence-Aligned European Patent Corpus. In: Proceedings of the 15th Conference of the European Association for Machine Translation, Leuven, Belgium, pp. 177–184 (2011)
6. Utiyama, M., Isahara, H.: A Japanese-English Patent Parallel Corpus. In: MT Summit XI, pp. 475–482 (2007)
7. Lu, B., Tsou, B.K., Tao, J., Kwong, O.Y., Zhu, J.: Mining Large-scale Parallel Corpora from Multilingual Patents: An English-Chinese example and its application to SMT. In: Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing, pp. 79–86. Chinese Information Processing Society of China, Beijing (2010)
8. Haegeman, L.M.V., Guéron, J.: English grammar: a generative perspective. Blackwell, Oxford (1999)
9. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening Ontologies with DOLCE. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 166–181. Springer, Heidelberg (2002)