Insight to Hyponymy Lexical Relation Extraction in the Patent Genre Versus Other Text Genres*

Linda Andersson, Mihai Lupu, João Pallotti, Florina Piroi, Allan Hanbury, Andreas Rauber
Vienna University of Technology
Information and Software Engineering Group (IFS)
Favoritenstrasse 9-11, 1040 Vienna, Austria
surname@ifs.tuwien.ac.at

ABSTRACT

Due to the large amount of available patent data, it is no longer feasible for industry actors to manually create their own terminology lists and ontologies. Furthermore, domain specific thesauruses are rarely accessible to the research community. In this paper we present extraction of hyponymy lexical relations conducted on patent text using lexico-syntactic patterns. We explore the lexico-syntactic patterns. Since this kind of extraction involves Natural Language Processing we also compare the extractions made with and without domain adaptation of the extraction pipeline. We also deployed our modified extraction method to other text genres in order to demonstrate the method's portability to other text domains. From our study we conclude that the lexico-syntactic patterns are portable to domain specific text genre such as the patent genre. We observed that general Natural Language Processing tools, when not adapted to the patent genre, reduce the amount of correct hyponymy lexical relation extractions and increase the number of incomplete extractions. This was also observed in other domain specific text genres.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: [linguistic processing]

Keywords

Patent Text Mining, Natural Language Processing, Ontology

1. INTRODUCTION

One of the first tasks of a patent examiner when given a new patent application is to identify essential patent aspects and extract terms, which later can be used in the search query session.

When conducting Prior Art Search it is essential to find different aspects of a patent? Each aspect can be divided into term pairs consisting of a general term and a specific term. [1]

This task requires both domain knowledge and access to technical terminology (both explicit and implicit knowledge). However, previous studies in the patent genre have observed that patent writers intentionally use entirely different word combinations to re-create a "concept", which increases the vocabulary mismatch issue [5]; and thereby make commercial technical terminology dictionaries such as EuroTermBank[1] and IATE[2] less re-usable [17].

In this paper we explore hyponymy relation extraction from the collection itself using lexico-syntactic patterns defined in [13]. With the variation in concept formulations, where paraphrasing of existing concepts is generally applied, a support tool such as a thesaurus or an ontology based on automatic extraction of lexical relations from the patent genre will be an usable search aid. Automatic ontology population consists of several steps, normalization of data, tokenization, Part of Speech (PoS) tagging, etc. However, the problem of using standard Natural Language Processing (NLP) tools is that the source data and the target data do not have the same feature distribution, this being a pre-requisite for their correct use [26]. Too many unseen events will decrease the performance of broad coverage NLP tools. In order to reduce the gap between source and target data several studies involving patent domain adaptation of broad coverage NLP tools have been investigated [16, 10, 2, 11, 23, 8]. The focus of these adaptations have been either on reducing the sentence length or increasing the lexicon. Only [2] and [8] have target adaptations incorporating domain information about the noun phrases' (NP) syntactic distributions. In this paper, we re-use the heuristic rules presented in [2].

The objectives of this study are:

1. to examine if it is possible to extract hyponymy lexical relations using the general lexico-syntactic patterns defined in [5];
2. to verify if the heuristic domain adaptation rules deployed in the extraction pipeline improve the candidate extractions;
3. to examine the portability of our modified extractor method, developed for the patent text genre, to other domain specific genres;
4. to examine if it is possible to simplify the evaluation process of hyponymy relation extraction.

The remainder of this paper is organized as follows. We first present some related work and terminology in Section 2. In Section 3 we present our experimental set up. In Section 4 we report our general results. Section 5 presents our conclusion and future work.

Published at Ceur-ws.org
Proceedings of the First International Workshop on Patent Mining and Its Applications (IPAMIN) 2014. Hildesheim. Oct. 7th. 2014. At KONVENS'14, October 8–10, 2014, Hildesheim, Germany.

[1] http://project.eurotermbank.com/
[2] http://iate.europa.eu/

2. RELATED WORK

In the Information Retrieval (IR) community, the patent retrieval research has focused mainly on improvements and method developments within systems for supporting patent experts in the process of Prior Art search. Less research attention has been given to other type of resources that support the patent examiner in the information process activities.

2.1 Terminology Effect on NLP

Before we can re-use NLP tools in text genres with high density of scientific terminology and new words, we need to understand the word formation process of the English language. The most productive word formation in English is affixation i.e. adding prefixes or suffixes to a root [6]. The suffixes '-ing' and '-ed' are especially problematic for NLP applications because when they are added to verbs, the new formed word may be a noun, an adjective or remain a verb (as in sentence 8, Figure 1 in the Appendix).

One of the major mechanisms of word formation is the morphological composite, which allows the formation of compound nouns out of two nouns (e.g. floppy disk, air flow) [18], and thereby creating a Multi Word Unit (MWU). It has been observed that in the technical jargon a heavy use of noun compounds constitutes the majority of scientific terminologies [14]. The compounding strategy causes not only unseen events on word level with new orthographical units, it also generates a diversity of syntactic structures among noun phrases, which is problematic for NLP tools [10, 24]. Furthermore, many NLP applications have chosen to overlook MWUs due to their complexity and flexible nature [4].

NPs can consist of single tokens, or can as well be as long and complex as any other occurring phrases in a sentence [15]. The NPs have an internal structure that dictates where additional elements can occur in relation to the head noun (e.g. pre- and post-modifier). There is a range of elements that can take the pre-modifier role in an NP but adjectives are the most typical pre-modifiers. In hyponymy lexical relation extraction, adjectives have a semantic significance, since the adjective modifiers could be considered a hyponym to the head noun [7]. For example, 'apple juice' is a valid hyponym to 'juice', but only in this combination since the modifier 'apple' specifies the head 'juice' [6]. The post-modifier construction is more complex, since a head noun can be post-modified by both phrases and clauses.

One central concept when analyzing NPs is to define its head [24]. The head in an NP has a supreme importance, as is the central part of the noun (e.g."the poet Burns", "Burns the poet") [15]. When a NP contains a prepositional phrase the traditional linguists promote the proper name (e.g. "the city of Stockholm") or the NP followed after the preposition (e.g. "a group of DNA strings") as the main head noun, since the NP after the preposition tends to have the highest degree of lexicalization [6, 24, 15]. However, what should be identified as the head noun in an NP is not straight forward [24]. Moreover, in [10] it was observed that the syntactic parsers right-headed bias caused problems during the analysis step of the patent sentences, thereby yielding erroneous analyzes.

2.2 Patent Text Effects in NLP

Patents are semi-structured documents which offer many different applications for text mining [3]. In patent documents, abstract and non-standard terminology is used to avoid narrowing the scope of the invention, unlike the style of other genres like newspapers and scientific articles [21]. Moreover, the vocabulary varies over time with terms such as "LP" and "water closet" being regarded as instances of obsolescence [12]. This type of discourse characteristic makes the patent text mining task more challenging. Many Patent Retrieval studies have tried to address different patent search problems by applying linguistic knowledge, using broad coverage NLP tools. However, as the generic NLP tools are not trained on the patent domain they experience problems with parsing long and complex NPs [10, 8]. There have been several studies focusing on reducing the gap between the source and target data, the focus being placed mainly on sentence reduction [11], on lexicon increase [16], or on both [23]. However, just increasing the lexical coverage or decomposing sentences will not solve the problem, since token coverage and sentence length are only part of the problem. [28] concluded that, since there is no significant difference between the general English and the English used in the patent discourse, on single token coverage, the technical terminology is more likely present in multi-word constructions consisting of complex NPs. Information about NPs' syntactic distribution has only been deployed in [2, 8], in order to improve the NLP analysis. In [8] a hierarchical chunker was designed to fit the syntactic structure of the patent sentence, targeting embedded NPs, while in [2] heuristic rules addressing the most common observed errors made by the NLP tools were used as a post correcting filter.

2.3 Ontology Population

Automatic ontology population relates to the methods used in Information Extraction (IE) as the general purpose is to extract predefined relations from text, hence referred to ontology based information extraction (OBIE) [19]. There are several applications where OBIE is used to enhance domain knowledge, to create a customized ontology, and in rich existing ontologies. OBIE techniques consist of identifying named entities (NE), technical terms, or relations. The OBIE process consists of several steps, data normalization, tokenization, PoS tagging, etc., thereafter following the recognition steps like gazetteers combined with rule-based grammars, ontology design pattern (ODP), pattern slots identifications such as lexico-syntactic pattern (LSP). Different techniques for hyponymy lexical relation extraction have been explored âĂŞ many of them depending on pre-encoded knowledge such as domain ontologies and machine readable dictionaries [9]. In order to avoid the need of pre-existing domain knowledge and remain independent of the sub-language one option is to use generic LSPs for hyponymy lexical relation extraction. [13] proposed a method to extract hyponymy lexical relations based on five LSPs, see Table 1.

Table 1: Sentence examples to each lexical syntactic pattern

Example sentences		LSP
1	... work such author as Herrick, Goldsmith, and Shakespeare	such NP as {NP, }* {(or|and)} NP
2	Even then, we would trail behind other European Community member, such as Germany, France and Italy	
3	Bruises, wounds, broken bones or other injuries	NP{, NP}*{,} or other NP
4	Temples, treasuries, and other important civic buildings	NP{, NP}*{,} and other NP
5	All common-law countries, including Canada and England	NP{,} including {NP,}* {or|and} NP
6	... most European countries, especially France, England, and Spain	NP{,} especially {NP,}* {or|and} NP

There are several issues related to extracting relations from a raw text based on LSPs. For instance, the LSP examples 2, 5 and 6 in Table 1 are not clear cases of hyponymy lexical relations, as in

'domestic pets such as cats and dogs,' since in LSP 2 Germany, France and Italy are members of the European Community and in LSP 6 France, England and Spain are countries in Europe i.e. a part of the geographic content called Europe [20].

With a wider semantic definition of the hyponym property, we can include both 'part of' and 'member of' in the definition:

"... an expression A is a hyponym of an expression B iff the meaning of B is part of the meaning of A and A is subordinated of B. In addition to the meaning of B, the meaning of A must contain further specifications, rendering the meaning of A, the hyponym, more specific than the meaning of B. If A is a hyponym of B, B is called a hypernym of A." [18, p83]

Hearst's patterns, [13], give high precision but low recall, while ODP gives high recall and low precision [19]. In [13], LSP 1 was used to extract candidate relations from the Grolier's American Academic Encyclopaedia (8.6M words). In this study, 7,067 sentences match LSP 1 and 152 relations fit the restriction i.e. to contain an unmodified noun (or with just one modifier).

A common approach to evaluate hyponymy relation extractions is to use an existing ontology as a gold standard [9]. For instance, in [13] the assessment was conducted by looking up if the relation was found in WordNet. Out of 226 unique words, 180 words existed in the WordNet hierarchy, and 61 out of 106 relations already existed in the WordNet. However, since most of the terms in WordNet are unmodified nouns or nouns with a single modifier, using WordNet in the evaluation process of this study was not feasible.

In [5] the gold standard was created by using linguists, but this type of labeling task is both time-consuming and costly, which makes the approach feasible only for small gold standards. The annotators were asked to manually identify domain-specific terms, NEs, synonymy and hyponymy relationships between identified terms and NEs. The annotation task requires both linguistic knowledge, as well as, some domain specific knowledge.

The gold standard was used to evaluate automatic hyponymy relation extractions from technical corpora, in English and Dutch. The data consisted of dredging year reports and news articles from the financial domain. The data was enriched with PoS tagging and lemmas produced by the LeTs Preprocessing Toolkit. The LeTs Preprocessing toolkit was trained on similar data where the accuracy of the PoS tagger was 96.3%. The NE extractor only achieved a recall of 62.92% and a precision of 59.33% [27].

For the hyponymy lexical relations extraction, three different techniques were used: 1) a lexico-syntactic pattern model based on LSP in [13], 2) a distribution model using context cluster by an agglomerative clustering technique and 3) a morpho-syntactic model. The morpho-syntactic model is based on the head-modifier principle:

- Single-word NP, if lexical item L_0 is a suffice string of lexical item L_1, L_0 is a hypernym of L_1
- MWUs NP, if lexical item L_0 is the head of term of lexical item L_1, then L_0 is a hypernym of L_1
- NP + prepositional phrase, if lexical item L_0 is the first part of a term in L_1 containing a NP plus prepositions (EN: of, for, before, from, to, on), then L_0 is to be the hypernym of L_1.

[17] concluded that the pattern-based methods and especially the morpho-syntactic approach achieved good performance on the technical domain data, therefore demonstrating that the general purpose hypernym detection models are portable to other domain and user-specific data.

In [21], hyponymy relations were extracted from US and Japanese patent re-using LSP patterns in [13]. For English 3,898,000 and for Japanese 7,031,149 candidate hyponymy relations were identified. The alignment between the language pair was conducted via citation analysis; 2,635 pairs of English-Japanese hyponymy relations were manually evaluated. The best method obtained Recall of 79.4% and Precision of 77.5%.

3. OUR APPROACH

Our data sets consist of five different text genres: the Brown corpus[3] (henceforth Brown), the WO and EP patent documents of IREC (Patent)[4], the TREC test collection for Clinical Decision Support Track (MedIR)[5], the test collection for Mathematical retrieval provided by NTCIR (MathIR)[6], and the papers produced during the Conference and Labs of Evaluation forum[7] (CLEFpaper). In Table 2 we present the total amount of sentences fitting the LSPs per data and extraction methods.

Table 2: Sentences per LSPs, data collection and extraction method.

	Patent	MedIR	MathIR	CLEF paper	Brown
Domain Rules	92,702	1,643,254	48,922	3,698	762
Simple Rules	135,550	2,084,529	70,822	5,748	950
No Rules	135,946	2,252,056	73,472	6,164	944

Example sentences from each data sets are shown in the Appendix , Figure 1.

3.1 Method

For this experiment we applied exactly the same methodology to all 5 data sets. We used all of the LSP patterns in Table 1. For the NLP pipeline we enriched all data sets with PoS tags using the Stanford tagger – English-left3words-distisim.tagger model [25]. In order to allow more flexibility to the phrase boundary we chose to use the baseNP Chunker [22]. We defined three pipeline extraction methods:

1. No rules (NoRules) was used to modifying the NLP pipeline analyzes
2. Three rules (SimpleRules) addressing observed errors among sentence fitted the LSP patterns. The rules address different type of conjunction and commas issues. Rule i) NP [cat and dogs] changed to two NPs [cat] and [dog], ii) [cat or dogs] changed two NPs [cat] or [dog], iii) numerous listing with commas.

[3] http://www.hit.uib.no/icame/brown/bcm.html

[4] IREC, is the corrected version of the MAREC http://www.ifs.tuwien.ac.at/imp/marec.shtml

[5] http://www.trec-cds.org/2014.html

[6] http://ntcir-math.nii.ac.jp/

[7] http://www.clef-initiative.eu/publication/proceedings

3. Domain rules, (DomainRules) here we applied the simple rules (2) and the rules presented in [2].

Figure 1 in the appendix displays the difference between NoRules and DomainRules among the pairs of sentences (3,4), (5,6) and (7,8).

Table 4: Correct identified positive relations and NP boundaries in relation to sample and for the most dominant relation "A kind of"

Group:Linguist		DomainRules		NoRules		SimpleRules	
		hyper ok	hypo ok	hyper ok	hypo ok	hyper ok	hypo ok
Brown	A kind of	70%	78%	**71%**	***83%***	**71%**	***83%***
	Relations	**72%**	80%	70%	***84%***	69%	83%
MedIR	A kind of	84%	***96%***	84%	93%	**93%**	91%
	Relations	**87%**	**92%**	**87%**	***92%***	**87%**	88%
MathIR	A kind of	**85%**	78%	64%	64%	65%	***79%***
	Relations	**86%**	77%	68%	***82%***	68%	81%
CLEF paper	A kind of	71%	***90%***	**75%**	75%	71%	83%
	Relations	76%	***89%***	**77%**	88%	74%	84%
Patent	A kind of	**82%**	***92%***	76%	76%	79%	90%
	Relations	79%	***91%***	77%	90%	**80%**	90%

Table 5: Number of positive extraction in relation to all extraction made for each sample and method

Group:Linguist	DomainRules	NoRules	SimpleRules
Brown	39%	**40%**	**40%**
MedIR	52%	33%	**54%**
MathIR	44%	**66%**	33%
CLEFpaper	50%	47%	**56%**
Patent	64%	71%	**81%**

For the evaluation only a smaller set was sampled out (1,647 instances) for manual assessment, approximately 100 instances per data collection and method. One instance correspond to one relation extracted from a sentences, if there are several possible extraction in a single sentence, each extraction correspond to one instance (see figure 1 in the appendix). Therefore not exact 1,500 instances were evaluated since some sentences contain more than one instances. Due to the fact that there are very few people having the level of linguistic knowledge, as well as the domain specific knowledge required to conduct assessment, we decided upon a more generic evaluation schema. The assessors were divided into three groups: linguist, and expert and non-expert. The linguist has domain knowledge of the patent domain and the computer science domain.

For the evaluation task, we constructed a simple interface, see figure 1 in the appendix. The evaluation tool shows the original sentence and five definition of relations between L_0 and L_1; i) L_0 is a kind of L_1, ii) L_0 is a part of L_1, iii) L_0 is a member of L_1; iv) L_0 is in another relation with L_1, v) L_0 has no relation to L_1. For uncertainty of the assessor we added Cannot say anything about the two and for erroneous extraction we added The sentence makes no sense. Since the NP boundaries were not entirely correct identified for all extractions, we added a check box for wrong boundary (for L_0 and L_1). In the instruction for the evaluation task, a simple example and a domain example were given for all types of relations.

In order to find out how difficult the task was thought to be by the assessors, we asked each assessor to grade each relation from as scale 1 (very easy) to 5 (very difficult). Furthermore, since it was observed in [3] that web searches for many candidate phrases were required in order to understand their meaning, we gave the assessor the possibility to search for the concept via a web service. We aim to improve the evaluation tool and give better interactive support therefore this feedback information is valuable for us.

4. RESULTS

In Table 3 we present the evaluation result based upon the linguist assessor. We see that the NoRules method generates more candidate extractions compared to the other ones, with correct boundary identification. This fact puzzled us since our experience during the assessment indicated the opposite. For instance, a common error was deverbal nouns exclusion. This error especially decreased correct and complete extractions for the domain specific text genres when using NoRules. For instance, when the head noun is a deverbal noun, the PoS-tagger assigns the label verb instead of a noun (e.g. "ultrasonic/JJ welding/VBN" and "laser/NN welding/VBN", and compare sentences 7 and 8 in figure 1, appendix).

Our first assumption to this contradiction was that one of the rules in the DomainRule method, which unifies NPs with 'of'-construction, harmed the extractions. In example 1, the hypernym consists of an embedded NP with prepositional 'of'-construction modifying the head noun.

Example 1: Embedded NP 'of'-construction

> The novel conjugate molecules are provided for the manufacture of a medicament for gene therapy, apoptosis, or **for the treatment of diseases** such as *cancer, autoimmune diseases* or *infectious diseases.*

If we include the entire NP i.e. "the treatment of diseases" the hyponymy lexical relation becomes incorrect since "cancer", "autoimmune diseases" and "infectious diseases" are "diseases" and not "treatments". On the other hand, in sentence 5 (figure 1, appendix) the relation between the hypernym and hyponyms becomes incorrect since hyponyms constitute properties of the hyponym therefore the NP should be unified. In sentences 3 and 4 (figure 1, appendix) the unification of the NPs 'of'-construction is more doubtful for the hypernym where "potential risk factors" (sentence 3) compared to "the distribution of potential risk factors" (sentence 4) seems to be the better choice. However, one of the hyponyms is overlooked in sentence 3 but extracted in sentence 4 with the help of the domain rule unifying 'of'-construction NPs. When examining the outcome of the rule we found that 131 instances were considered correct (i.e. the NP with 'of'-construction should be unified) and only 44 instances were incorrect. The more likely reason for the NoRules more complete and correct identified hyponymy relation is that the NoRules generated more extractions compared to DomainRules which has a more strict extraction rule schema.

Table 4 shows the percentage of the most dominant relation "a Kind Of" and all positive relations ("a Kind Of", "a Part Of", "a Member Of", "another Relation") for each method and data set. The preferred hypernym rule is the DomainRules method regardless of data set. For hyponyms, the result is more inconclusive since several methods ended up having the same percentages. For the "a Kind Of" relation the preferred method is either SimpleRules or NoRules as seen in Table 4.

Table 5 displays the percentage of all examined sentences matching the LSP patterns where a positive and correct extraction was identified. For three out of five data sets the method SimpleRules was

Table 3: The total amount of correct identified relation and NP boundaries

Group: Linguist	DomainRules			NoRules			SimpleRules		
	hyper ok	hypo ok	Total	hyper ok	hypo ok	total	hyper ok	hypo ok	total
Brown	74	82	103	**94**	**113**	135	95	114	137
MedIR	110	116	126	**150**	**159**	172	142	144	163
MathIR	83	74	96	**84**	**101**	123	70	83	103
CLEFpaper	70	82	92	**99**	**113**	129	86	98	117
Patent	109	125	138	**147**	**172**	191	150	169	188

Table 6: Inter-annotator agreement between assessment groups

	MathIR		Brown	CLEFpaper
	Linguist vs Expert	Linguist vs None Linguist	Linguist vs None Linguist	Linguist+Domain knowledge vs Expert
Relations	85%	81%	83%	88%
No relation	68%	72%	72%	75%
Cannot tell	86%	77%	83%	89%
Makes no sense	90%	89%	80%	93%
hypernymBoundaryWrong	64%	67%	83%	67%
hyponymBoundaryWrong	62%	67%	85%	82%

preferred.

In order to examine the simplification of the evaluation process, we computed inter-annotation agreements between the three groups: expert, linguist and non-expert. The inter-annotation agreement for identifying relations ranges between 81% and 88% (Table 6), regardless of the group comparisons for Brown and for the scientific paper data sets. Similar agreement values were found for the patent and medical text domain. The inter-annotation agreement decreases for wrong NP boundary identifications, which can be explained by that fact that it requires linguistic schooling to correctly identify NPs.

5. CONCLUSIONS

We conclude the following:

- It is possible to re-use LSPs for hyponymy lexical relation extractions. We thereby confirm the observation made in [1] that the LSP method for relation extraction is portable to different text genres.
- We also confirm that for domain specific text genre, such as patent or medical genres, at least for the hypernyms modification of NLP tools is required. For detecting hyponyms the additional rules were less successful. On the other hand, as seen in sentences 7 and 8 (figure 1, appendix) the rules addressing deverbal nouns make it possible to extract more correct instances.
- The simplified process of evaluating hyponymy lexical relations extractions using non-linguists and non-experts is on an acceptable inter-annotation agreement level. However, more information regarding the identification of NP boundaries should be added in future evaluation guidelines.
- In the future we will explore machine learning algorithms to select which extraction method should be used for a specific relation, instance and data collection. The additional modifying the NLP pipeline need further examination, since it becomes contra productive for some instance but improve for others. Furthermore, we also want to examine additional patterns exploring similarity between the internal structures of NPs, as described in [1].

In the future we will explore machine learning algorithms to select which extraction method should be used for a specific relation, instance and data collection. The additional modifying the NLP pipeline need further examination, since it becomes contra productive for some instance but improve for others. Furthermore, we also want to examine additional patterns exploring similarity between the internal structures of NPs, as described in [1].

6. ACKNOWLEDGMENTS

This research was partly funded by the Austrian Science Fund (FWF) projects P25905-N23 (ADmIRE) and I1094-N23 (MUCKE).

7. REFERENCES

[1] A. Adams. Personal correspondence. PatOlympics, 2011, Vienna, 2011.

[2] L. Andersson, M. Lupu, and A. Hanbury. Domain adaptation of general natural language processing tools for a patent claim visualization system. In M. Lupu, E. Kanoulas, and F. Loizides, editors, *Multidisciplinary Information Retrieval*, volume 8201 of *Lecture Notes in Computer Science*, pages 70–82. Springer Berlin Heidelberg, 2013.

[3] P. Anick, M. Verhagen, and J. Pustejovsky. Identification of multiword expressions in the brwac. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of LREC-2014*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[4] V. Arranz, J. Atserias, and M. Castillo. Multiwords and word sense disambiguation. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 3406 of *Lecture Notes in Computer Science*, pages 250–262. Springer Berlin Heidelberg, 2005.

[5] K. H. Atkinson. Toward a more rational patent search paradigm. In *Proceedings of the 1st ACM Workshop on Patent Information Retrieval*, PaIR '08, pages 37–40, New York, NY, USA, 2008. ACM.

[6] K. Ballard. *The Frameworks of English*. Palgrave Macmillan, 2007.

[7] W. Bosma and P. Vossen. Bootstrapping language neutral term extraction. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias,

editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[8] N. Bouayad-Agha, A. Burga, G. Casamayor, J. Codina, R. Nazar, and L. Wanner. An exercise in reuse of resources: Adapting general discourse coreference resolution for detecting lexical chains in patent documentation. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of LREC-2014)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[9] P. Cimiano. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[10] E. D'hondt, S. Verberne, C. H. A. Koster, and L. Boves. Text representations for patent classification. *Computational Linguistics*, 39(3):755–775, 2013.

[11] G. Ferraro. *Towards deep content extraction from specialized discourse: the case of verbal relations in patent claims*. PhD thesis, Universitat Pompeu Fabra, 2012.

[12] C. G. Harris, R. Arens, and P. Srinivasan. Using classification code hierarchies for patent prior art searches. In M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, editors, *Current Challenges in Patent Information Retrieval*, volume 29 of *The Information Retrieval Series*, pages 287–304. Springer Berlin Heidelberg, 2011.

[13] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[14] J. S. Justeson and S. M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27, 3 1995.

[15] E. Keizer. *The English Noun Phrase: The nature of linguistic categorization*. Cambridge University Press, 2010.

[16] C. H. Koster and J. G. Beney. Phrase-based document categorization revisited. In *Proceedings of the 2Nd International Workshop on Patent Information Retrieval*, PaIR '09, pages 49–56, New York, NY, USA, 2009. ACM.

[17] E. Lefever, M. V. de Kauter, and V. Hoste. Evaluation of automatic hypernym extraction from technical corpora in english and dutch. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of LREC-2014*, pages 490–497, 2014.

[18] S. Löbner. *Understanding Semantics*. Oxford University Press, New York, 2002.

[19] D. Maynard, Y. Li, and W. Peters. Nlp techniques for term extraction and ontology population. In *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap Between Text and Knowledge*, pages 107–127, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press.

[20] V. Mititelu. Automatic Extraction of Patterns Displaying Hyponym-Hypernym Co-Occurrence from Corpora. In *Proceedings of the 1st CESCL, Budapest, Hungary*, 2006.

[21] H. Nanba, S. Mayumi, and T. Takezawa. Automatic construction of a bilingual thesaurus using citation analysis. In *Proceedings of the 4th Workshop on Patent Information Retrieval*, PaIR '11, pages 25–30, New York, NY, USA, 2011. ACM.

[22] L. A. R. and M. P. M. Text chunking using transformation-based learning. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, volume 11 of *Text, Speech and Language Technology*, pages 157–176. Springer Netherlands, 1999.

[23] S. Sheremetyeva. Natural language analysis of patent claims. In *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing - Volume 20*, PATENT '03, pages 66–73, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[24] K. Spärck Jones. Compound Noun Interpretation Problems. In F. Fallside and W. A. Woods, editors, *Computer Speech Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1983.

[25] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[26] J. Turmo, A. Ageno, and N. Català. Adaptive information extraction. *ACM Comput. Surv.*, 38(2), July 2006.

[27] M. van de Kauter, G. Coorman, E. Lefever, B. Desmet, L. Macken, and V. Hoste. LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal*, 3:103–120, 12/2013 2013.

[28] S. Verberne, C. H. A. Koster, and N. Oostdijk. Quantifying the challenges in parsing patent claims. In *In Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval (AsPIRe 2010*, pages 14–21, 2010.

Collecton	#	Sentence	Correct	Example
CLEF paper	1	For/IN now/RB ,/, [we/PRP] do/VBP not/RB treat/VB **[compounds/NNS]** ,/, [proper/JJ nouns/NNS] ,/, [acronyms/NNS] or/CC [other/JJ entities/NNS] ./.	3	entities ->proper nouns
Brown	2	[Residential/NNP building/NN] consists/VBZ of/IN [houses/NNS] ,/, [apartments/NNS] ,/, [hotels/NNS] ,/, [dormitories/NNS] and/CC [other/JJ buildings/NNS] designed/VBN for/IN [shelter/NN] ./.	3	buildings ->hotels
MedIR No Rules	3	[The/DT distributions/NNS] of/IN [potential/JJ risk/NN factors/NNS] ,/, such/JJ as/IN [first-degree/JJ family/NN history/NN] of/IN [breast/NN cancer/NN] ,/, [body/NN mass/NN index/NN] ,/, [alcohol/NN consumption/NN] ,/, [menopausal/JJ status/NN] ,/, [parity/NN and/CC breastfeeding/NN] ,/, were/VBD similar/JJ in/IN [carriers/NNS and/CC noncarriers/NNS] of/IN [the/DT SULT1A1/NN] */SYM [2/CD allele/NN] ./.	3	potential risk factors ->body mass index
MedIR Domain Rules	4	**[The/DT distributions/NNS of/IN potential/JJ risk/NN factors/NNS]** ,/, such/JJ as/IN **[first-degree/JJ family/NN history/NN of/IN breast/NN cancer/NN]** ,/, [body/NN mass/NN index/NN] ,/, [alcohol/NN consumption/NN] ,/, [menopausal/JJ status/NN] ,/, [parity/NN and/CC breastfeeding/NN] ,/, were/VBD similar/JJ in/IN [carriers/NNS and/CC noncarriers/NNS] of/IN [the/DT SULT1A1/NN] */SYM [2/CD allele/NN] ./	4	*the distribution of potential risk factors* -> first-degree family history of brest cancer
MathIR-No rules	5	[We/PRP] have/VBP proven/VBN [the/DT basic/JJ properties/NNS] of/IN [these/DT special/JJ functions/NNS] ,/, such/JJ as/IN [recursion/NN relations/NNS] ,/, [orthogonality/NN] ,/, [differential/JJ equations/NNS] and/CC [the/DT like/JJ]	0	*ortogonality is a property of special function but not a special function by it self*
MathIR Domain Rules	6	[We/PRP] have/VBP proven/VBN [the/DT basic/JJ properties/NNS of/IN these/DT special/JJ functions/NNS] ,/, such/JJ as/IN [recursion/NN relations/NNS] ,/, [orthogonality/NN] ,/, [differential/JJ equations/NNS] and/CC [the/DT like/JJ]	3	basic property of special function -> recursion relations
Patent No Rules	7	Removing/VBG of/IN [the/DT solvent/JJ] of/IN [a/DT solvent/JJ coated/JJ layer/NN] may/MD be/VB effected/VBN by/IN [any/DT suitable/JJ conventional/JJ technique/NN] such/JJ as/IN [oven/NN] drying/VBG ,/, infrared/VBN [radiation/NN] drying/VBG ,/, [air/NN drying/VBG] and/CC [the/DT like/JJ]	1	suitable conventional technique -> air drying
Patent Domain Rules	8	Removing/VBG of/IN [the/DT solvent/JJ of/IN a/DT solvent/JJ coated/JJ layer/NN] may/MD be/VB effected/VBN by/IN [any/DT suitable/JJ conventional/JJ technique/NN] such/JJ as/IN **[oven/NN drying/VBG]** ,/, **[infrared/VBN radiation/NN drying/VBG]** ,/, [air/NN drying/VBG] and/CC [the/DT like/JJ]	3	suitable conventional technique -> infrared radiation drying

Figure 1: Sentences examples for the different data sets, with and without Domain Rules.

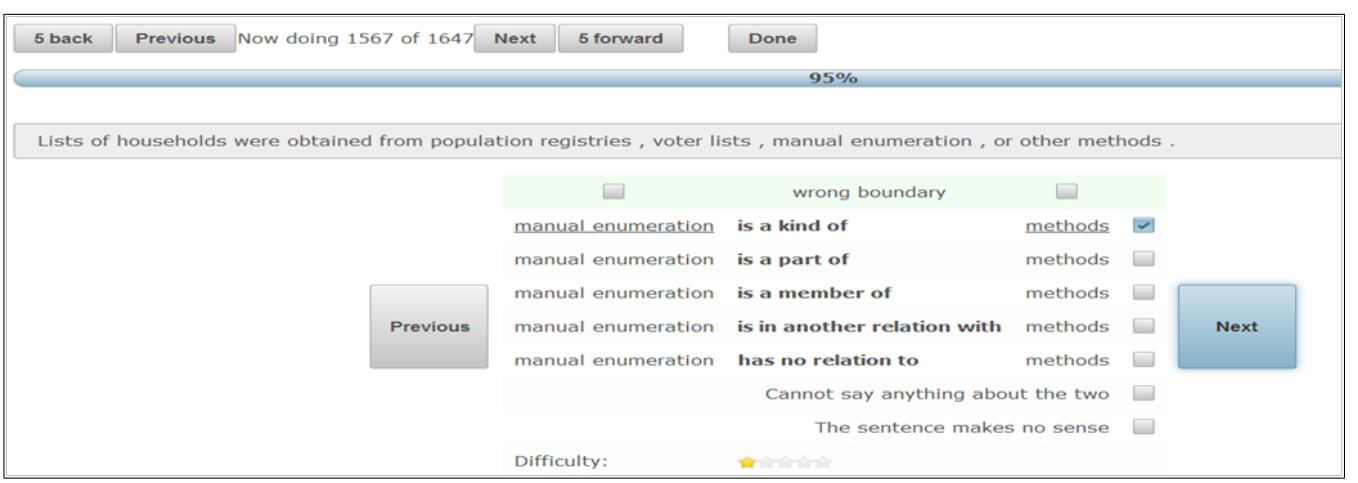

Figure 2: Evaluation tool interface.