

Lexical Issues of a Syntactic Approach to Interactive Patent Retrieval

Eva D'hondt

Centre for Language and Speech Technology,
Radboud University Nijmegen, The Netherlands
e.dhondt@let.ru.nl

Abstract

Patent retrieval is an information retrieval task that poses very specific characteristics and demands. Especially the need for high recall is very important to patent searchers. In the ongoing research project TM4IP, we aim to improve patent retrieval by developing an open-domain patent retrieval system based on linguistic knowledge. By using Dependency Triplets as index terms our system aims to improve precision and recall compared to keyword-based approaches. One of the cornerstones of a syntactic approach to Information Retrieval is normalisation. This paper describes some of the characteristics of the patent domain that influence lexical normalisation.

Keywords: Patent Retrieval, Natural Language Processing, Dependency Triples, Lexicon

1. INTRODUCTION

Over the last few decades, the increased access to patents and patent-related information has seriously changed the patent world. National patent offices' databases have become available online and the advent of machine translation and OCR technology has enabled patent searchers to widen their search areas and –at least in theory– the effectiveness of their search. *'However, the complexity and concern of getting it right is much higher too ... [among patent searchers, there is] a growing anxiety about missing something. [1]'* The economic repercussions may be vast: for example, a missed patent in a prior art search can lead to an infringement suit, which can cost millions of dollars. While in the past retrieval of patents and other forms of Intellectual Property was typically researched by the database community, more recently it has attracted the attention of the IR community. Patent retrieval has been the topic of workshops in [SIGIR 2000; ACL 2003; NTCIR3 2002; NTCIR4 2004] and this year the CLEF-IP 2009 track is directed exclusively at prior art search.

Patent retrieval is substantially different from ad hoc retrieval, because of the special characteristics of both the documents and the queries (as well as the goals of the searchers) that are involved in the search.

Depending on the specific purpose for undertaking the search, the end point of patent searches starting from the same query may be quite different. Prior art search, invalidity search, infringement watch or state-of-the-art search all have different goals and information needs and the relevance of the information in the set of found documents will be evaluated accordingly. This shows the need for a patent retrieval system to have a well-developed, interactive search component which can be modified to suit the goal and taste of the searcher.

When composing their search queries, patent searchers are very much concerned with the fact that they cannot afford to miss an important patent. I have dubbed this the 'total recall problem': The patent searcher is willing to lose precision in an effort to reach maximum recall. Expert users create long Boolean queries (comprising 5 to 30 terms) where each concept searched for is expressed by AND's and OR's of possible synonyms [3]. The resulting set of documents is then analysed document per document to judge their relevance. The exposure to alternative descriptions of some concept can lead to a careful rephrasing of the original query and so the

process is repeated until either the relevant information is found or the (known) list of variations is exhausted. This way of searching is very labour-intensive: some search sessions can take up to two months to locate 200 relevant patents [3]. There have been some attempts to facilitate this search process by automatically extending the search queries with synonyms or ontologically related terms, but such systems have not met great enthusiasm from the patent searchers community. As [5] point out:

'The professional searcher does not like the black box effect of intelligent engines, for example the automatic query expansion with synonyms, stemming, default use of the OR Boolean operator, etc. Whether a box is really black or just looks black because of absence of illumination (in this case knowledge of the linguistic algorithms used) can become the subject of philosophical speculations.'

While patent searchers realize that they need help to achieve the highest possible recall, they also demand total control and constant insight in precision and recall. Any retrieval system aimed at this specific task and public should be designed with these two concepts in mind: 'total recall' and 'total transparency towards the user'.

If we turn to the patent documents another set of problems appears:

Patent texts are not homogeneous, but consist of different sections¹ which use different styles. However, depending on the search goal the data relevant to (part of) the information need may be in any of these sections; so all of them must be taken into account. The claims section, for example, is legally bound to be one massive sentence. This is where the legal protection of the invention is determined, which results in even more legalese than usual.

Even in the more descriptive sections, the language of patent documents is still notoriously difficult to read. This is partly because of their grammatical structure: Patent texts often contain long, complex sentences with a lot of enumerations and ellipses. However, as argued by [8], in most patent texts, the grammar is correct and only a subset of the grammatical constructions in a language are used. This allows for an efficient analysis by a (specially developed) parser. Yet a greater challenge of processing patent texts, more specifically from an IR point of view, is the legal style in which patents are written. 'Patentese' combines the specific vocabulary of the domain to which the patent applies with very generic terms and expressions. (The latter is a tactic that is commonly used by patent lawyers who aim to obfuscate the method and specifics of the invention.) This huge variation in expressing concepts makes it very difficult to find all relevant documents pertaining to one's information need. As [1] puts it: *'In the patent world, words are shapeshifters, yet words are the brick and mortar of modern information retrieval.'*

The keyword-based systems that make up the majority of today's patent retrieval systems cannot look past this variation. A retrieval method based solely on surface forms encounters all sorts of problems: The difference between the noun and verb 'means' is undetectable for a pure keyword-based approach. A second problem encountered by these systems is the fact that they cannot deal with the morphological variation of word forms. Even in a keyword-based search which incorporates stemming, the fact that 'index' and 'indices' are basically the same word will not be noticed. Nor can keyword-based retrieval deal with the similarity between 'adhere' and 'adhesion' (nominalization). Because of the fundamental limitations of keyword search we have to go beyond the surface form. Therefore, we propose a syntactic approach to patent retrieval in which the syntactic (and correlated semantic) relations [10] between two words are the index terms. We use Dependency Triplets like [N:candle, SUBJ, V:burn] to represent different but equivalent expressions like 'the burning candle', 'the candle(s) burned', '(he was) burned by a candle', etc. By doing so we can abstract away to a level beyond surface realization, closer to concepts thus resolving problems of ambiguity due to morphological and syntactic variation.

¹title, abstract, description and claims section

My PhD research project is part of the TM4IP² project [4] which aims to develop an open-domain patent retrieval system based on linguistic knowledge. My research focuses on the linguistic concerns that arise while designing and implementing the parser and search module. Until now, the syntax and semantics of patent texts have received little attention from the linguistic point of view. While there are a multitude of problems and considerations that are relevant for linguistically informed patent search, the focus of this paper is more specifically on the lexical characteristics of patent texts which influence the process of transforming this type of text (genre) into usable/optimal index terms.

In section 2, I will briefly describe the state of the art in patent retrieval systems and give a brief overview of the system used in the TM4IP project. In section 3, I will zoom in on an experiment that I performed and discuss plans for future research.

2. BACKGROUND

2.1. Patent retrieval systems

The majority of the search engines used by the patent search community today are keyword-based, using a general-purpose text search engine. Some of them incorporate a query preprocessing module and allow for the use of wild cards, boolean compositionality and term weighting [12], query phrases, query expansion by using thesaurus [12], proximity search¹itemicropat, etc. For an overview of the three biggest commercial systems, see [11]. The vector space model usually lies at the heart of these generic keyword-based approaches. Academic research has mainly focussed on the relative weighing of these terms [6] and on exploiting the patent document structure to boost retrieval efficiency [6]. Over the last few years the first semantically based patent retrieval systems have been introduced, in particular the Patent-café search engine and IPCenturys DECOPA search engine.

The only method that comes close to our syntactic approach is [9] who uses deep linguistic analysis in the form of predicate-argument analysis (implying semantic role labelling) to improve readability. Her system is the first step in a suggested patent summarization method. A closely related system is the PATexpert system [2], a content-oriented system that aims to look past the surface forms and uses –amongst other technology like image retrieval, etc– semantic web technology to give direct access to the content of the patent. Due to the highly specialised ontologies used, the PATexpert system is currently (2008) focused on two domains: optical recording devices and machine tools. In TM4IP we want to avoid this dependence on elaborate ontologies and have opted for an open-domain system.

2.2. Introducing the PHASAR system in the TM4IP project

The basic unit in our system is the Dependency Triplet (DT). A DT is similar to a syntactic phrase in that it is a grammatical part of the sentence and –at least in part– identified according to linguistic criteria. The use of syntactic phrases in Information Retrieval is based on the assumption that words in a text that have a syntactic relationship often have a related semantic relationship [10]. A DT is a pair of (lemmatized) words together with their syntactic relation, e.g. [N:current, ATTR, A:electrical]. PHASAR's DT framework is based on the principle of aboutness.

Our system is made up of two main components:

a) AEGIR (Accurate English Grammar for Information Retrieval), a hybrid dependency parser, which aims to accurately parse complicated technical English texts for IR purposes. AEGIR combines a broad-coverage, handcrafted rule-based grammar with a transduction process to (a) find the best parse and (b) transduce this parse to DTs while processing large raw corpora in the patent domain. Then, the information about the frequency of DTs and lexical items is incorporated into the parser, thereby guiding the parsing process as it analyses new text. For example, the correct parse of the famous example John hit a man with a telescope could easily

² 'Text Mining for Intellectual Property.' For more information on the project, visit <http://www.phasar.cs.ru.nl/TM4IP.html>

be decided given the information that [V:hit, PREPwith, N:telescope] occurs 2 times in the entire (training) corpus, while [N:man, PREPwith, N:telescope] occurs 220 times.

b) PHASAR (Phrase-based Accurate Search And Retrieval) is a search engine which uses DTs as index terms. In the search module, the searcher can phrase queries in a semi-natural way to fit the index terms as closely as possible. One has complete control over the search and search result and can interactively generalize the query or make it more specific. Query generalization can be achieved by either joining multiple terms using the OR operator, or by using one of the built-in thesauri for selecting a semantic term type. A query can be made more specific, by adding more terms in the query slots or by setting a context from which the results have to be retrieved.

In our system, the normalisation needed to create effective index terms is taken care of by different parts of the system.

- **Grammatical normalisation** is built into the parser, which has a second transduction step that translates a dependency graph into a set of DTs. Special care has been taken to reach the highest grade of normalisation possible, for example by splitting up hyphenated forms such as 'man-made lake' into [N:man, SUBJ, V:make] [V:make, OBJ, N:lake], to map nominalisations onto the relevant verbs, to map equivalent grammatical structures onto each other, e.g. 'thumb movement' onto 'movement of thumb'.
- **Morphological normalisation** is realised partly through the lexicon and partly through the parser. Spelling variation and morphological variations such as singular-plural, tense or mood are handled by abstracting to a lemma (present in the lexicon). For those forms that have no entry in the lexicon, robust recognition rules are incorporated in the system.
- **Semantic normalisation** is (partly) realised by using the DTs themselves, which provide a disambiguating context (achieving higher precision). E.g. tree bark versus a dog that barks.
- **Lexical normalisation** is (partly) realised by the use of (several) thesauri which can be accessed by the user during a search and analysis phase.

3. CONSIDERATIONS AND POSSIBLE RESEARCH ISSUES

As explained in the previous section, problems of syntactic, morphologic and -to some extent- semantic variation are dealt with by the parser. But one problem still stands in the way of maximum recall: lexical variation. While being a big problem in any IR task, there are some characteristics of patent texts that make lexical variation even more difficult to deal with in the patent domain.

3.1. Diversity in the patent domain

The notion of the 'patent domain' is a deceptive one. The so-called patent domain actually consists of hundreds of highly technical and specialised subdomains, each with its own very specific terminology and ontologies. A patent document may concern anything from a gene extraction method, a business method or chemical compound to a particular design of a doll house.

With such a wide variety of specific subdomains, finding and extending reliable lexical resources is very difficult. Yet, ontologies and lexica are very important for all parts of our system. They are not only used for query extension or classification tasks, but the lexicon is the very basis of our parser system. When the parser encounters a word in the text that it cannot find in its lexicon, it can either ignore it, thus -unlike the bag-of-words approach- creating 'gaps' in the information processing of the text, or use a series of robust recognition rules to make an informed guess as to the relation of this lexical item compared to the other item in the dependency triplet. The latter approach is more error prone and often does not provide the correct POS information or lemma, thus inserting noise into the set of index triples.

3.2. Lexical and semantic variation

Even if we could have access to perfectly-constructed and complete ontologies, a second defining characteristic of the patent domain would still pose a serious problem: semantic variation or the possibility of one expression to denote several concepts. The patent domain exist through the

collaboration of (tens of) thousand of writers and as there is no-agreed upon vocabulary (like for example in the UMLS), every inventor has the right to use a word as he or she sees fit. For example, a multiword term such as 'sound emitting device' can denote everything from a car horn or headphones to the so-called Mosquito Device, a device which is designed to drive young troublemakers away from a problem area. This makes it very difficult to perform an effective keyword search/direct term-matching. To deal with this problem we need to disambiguate the meaning of the words that are encountered in the patent text.

Closely related to semantic variation is the problem of lexical variation: the same concept can be expressed in different patent texts with different lexical items, e.g. 'spring', 'wire of coiled steel', 'means compressible along an axis', etc. All these synonymous expressions should be identified as such and linked to each other by means of a thesaurus or ontology.

3.3. Vague terms

A third characteristic of patent texts is the use of vague terms and expressions. Terms such as 'sound emitting device', 'means compressible along an axis' or 'apparatus for preparing and dispensing whipped beverages' are used instead of 'loudspeaker', 'spring' or 'drink dispenser'. As mentioned above, this kind of lexical variation complicates direct term-matching. These terms are invented and defined ad hoc by the patent writers and will generally not end up in any dictionary or lexicon. Exceptions are terms whose acronyms have become common words, like LED (light emitting device) or LCD (liquid crystal display). Our system -at present- cannot catch these terms and will normalize them into separate DTs, essentially treating the information as if it would occur in a regular sentence instead of appearing as a term.³

While this syntactic normalisation usually improves recall [7], treating the information in sentences and inside terms in the same way has disadvantages for lexical normalisation. If a patent searcher is looking for all patents concerning headphones, he will also be interested in an obscure patent concerning a 'sound emitting device'. Instead of only splitting the information up in separate DTs, the system should be able to use these terms for the lexical normalisation process as well by linking them to known terms. This is equivalent to expanding the systems ontology with these ad-hoc synonyms. This requires a two-step approach: correctly identifying and extracting vague terms and linking them to an existing ontology.

Preliminary analysis of 16,000 patents in the engineering domain revealed that vague terms like the examples above are headed by so-called general-purpose words⁴. These words are semantically light: they occur quite often in the corpus (see appendix 2 for the occurrence of general-purpose terms in the 20 most frequent terms) but are, by themselves, not very informative. Their meaning is rather abstract and they are always accompanied by a verb or an adjective derived from a verb that specifies their function. I created a list of general-purpose words by looking at the 200 most frequent words and selecting the words which fit the criterion of 'whilst being an instrument or method the word does not contain an intrinsic expression of its function'. This was done by 2 persons with an inter-annotator agreement of 67%.⁵ The resulting list can be found in appendix 1. I then looked at the noun phrases which were headed by these words.

³The three examples given above would respectively become [N:device, SUBJ, V:emit], [V:emit, OBJ, N:sound]; [V:compress, OBJ, N:means], [V:compress, PREP:along, N:axis] and [N:apparatus, SUBJ, V:prepare], [N:apparatus, SUBJ, V:dispense], [V:prepare, OBJ, N:beverage], [V:dispense, OBJ, N:beverage], [V:whip, OBJ, N:beverage]. In this paper I do not discuss the more typical compound terms (a noun phrase in which the adjective or noun is attributively connected to the head noun), e.g. thermotherapeutic apparatus or liquid crystal display device. While these compounds are equally important in the search for lexical normalisation and will receive a great deal of attention in my research project, they are less specific for the particular difficulties of the patent domain than the 'vague terms'.

⁴A general-purpose word can be defined as noun that describes an article or instrument whose function is not expressed by the word itself but through the context words. For example, 'apparatus' in the term 'occupant sensing apparatus' does not express its function, while the word 'sensor' does.

⁵Inter-annotator agreement was calculated by counting each annotation of person 2 as a match or nonmatch value to the annotations of person 1 and then calculating the ratio of matches to the total number of annotations.

Manual analysis of the noun phrases showed that these vague terms usually follow one of these syntactic formats:

- Noun Phrase – Present Participle – General-Purpose Noun, e.g. 'cover folding device' or 'digital video/audio recording and reproducing apparatus'
- General-Purpose Noun – Adjective or Adjectival Phrase, e.g. 'means compressible along an axis',
- General-Purpose Noun – for/of – Present Participle Noun Phrase, e.g. 'device for making sandwiches', 'apparatus for dispensing medicine', 'a means and method for implanting bioprosthetic material'.

In a few patents small variations occurred within the vague terms. For example, a patent describing a 'cover folding device' would sometimes use the term 'top cover folding device' instead. These variations are a sign of the ad-hoc status of these terms.

Analysis of their distribution in the patent documents shows that these vague terms appear frequently in specific places such as the title, the abstract and the claims section. The description section of the document deals with the parts and components of the system, so naturally such general terms do not occur very frequently in this section. What is interesting, however, is the low frequency of these terms in the prior art descriptions. While one would expect relatively high frequency as they are describing similar concepts patent writers tend to opt for more concrete terms to denote the same concept. For example, the patent discussing an 'occupant sensing apparatus' would use the term 'sensor' in the prior art description, but not in the rest of the document.

The next step of my project will be to design a system that can a) automatically extract vague terms and b) link them to existing ontologies. One of the challenges of automatic term extraction will be to decide which words belong to the n-gram. As we are looking for relatively big terms (more than 3 words) the complexity of the task increases enormously. Next to our knowledge of the general-purpose terms list and the syntactic templates, one characteristic of patent documents that may help solve this problem is the patent writer's need to avoid ambiguity. As mentioned above there is a great deal of intertextual variation in the expression of concepts even within one domain. Yet here is almost no lexical variation in the patent texts themselves. Since failure to describe an invention or claim clearly and unambiguously can result in rejection during prosecution [1], the patent lawyer will not risk any misunderstanding and thus will always refer to a concept that has already been introduced in the text by means of clear anaphoric elements or will just repeat the term in full. In due time we will develop an anaphora resolution module, but for now the system can only look at those instances where the whole vague term was repeated. Using the linguistic knowledge of the parser to detect the three syntactic templates described above we can detect the potential vague terms. If such a term appears in the title, abstract and/or frequently⁶ in the claims section of the patent document, the system would classify it as a 'vague term'.

In subsequent work, the automatically extracted vague terms should be used to expand existing ontologies. We would like to know if a vague term could be a synonym for a more concrete term that is part of the ontology. As mentioned above, in the prior art description the patent writer often uses more concrete terms than in the rest of the patent to describe the same or a very similar concept or invention⁷. My hypothesis is that the verb in the vague term is a very good indicator of the desired (more concrete) term. It describes the function of the vague term and often occurs (almost) literally in the concrete term. For example, an 'occupant *sensing* apparatus' could easily be linked to 'occupant *sensor*' (using the nominalisation database NOMLEX). I will investigate if focusing on the term (and its synonyms in WordNet) can help identify less obvious concrete terms in the prior art section.

⁶That is, appear more often than a not yet specified cut-off ratio, normalised to the length of the claims section and abstract.

⁷For example, a patent concerning a 'device for cooling an infant's brain' in which this term was used very consistently used 'cooling cap' in the prior art section.

4. CONCLUSION AND DISCUSSION

In this paper I presented an overview of the characteristics of patent retrieval and discussed some of the lexical issues that are particular to the patent domain. These issues were described in the context of the TM4IP project, in which we develop a patent retrieval system based on a syntactic approach. The design behind the PHASAR search engine aims to get very high recall by achieving syntactic and morphological normalisation. We are currently looking for ways to extend our approach by adding extra lexical normalisation.

As was shown in section 3 lexical variation is an important problem in patent retrieval. The specific requirements of the patent searchers (i.e. 'total recall' and 'total transparency') demand a novel approach that combines both an automatic extraction of potential new terms and an interactive component allowing the patent searcher to keep control over the search terms in the query. In section 3.3, I sketched an initial implementation but this is by no means complete. Hence, I am open to any advice or comments on the current approach, or on alternatives and extensions of the approach. Specifically, there are a number of open issues that I would like to get feedback on:

- If vague terms can be identified, there still is the challenge of attaching them at the correct positions in the ontologies. What would be a good way to automatically extend existing ontologies?
- Determining which is the correct synonym will prove to be a major challenge. What do you imagine would be the major pitfalls?
- For those who have a background in patent retrieval: there is a common feeling that different sections of patent texts contain very different information, yet precious little has been published on this subject. Could anyone give me some pointers?

REFERENCES

- [1] Atkinson K. (2008) Toward a more rational patent search paradigm. In *Proceedings of the 1st ACM Workshop on Patent Information Retrieval*, (Napa Valley, California, USA, October 30 - 30, 2008). PaIR '08. ACM, New York, NY, 37-40.
- [2] Escorsa E, Giereth M, Kompatsiaris Y, Papadopoulos S, Pianta E, Piella G, Puhlmann I, Rao G, Rotard M, Schoester P, Serafini L and Zervaki V. (2008) Towards content-oriented patent document processing. In *World Patent Information* 30(1):21:33.
- [3] Homan H. (2004) Making the Case for Patent Searchers. In *Searcher*, 12(3).
- [4] Koster C, Oostdijk N, Verberne S and D'hondt E. (2009) Challenges in Professional Search with PHASAR. In *Proceedings of DIR 2009*, pp. 101-102
- [5] Krier M, Zacca F. (2002) Automatic categorisation applications at the European patent office. In *World Patent Information* 24(3):187-96.
- [6] Mase H, Matsubayashi T, Ogawa Y, Iwayama M, and Oshio T. (2005) Proposal of two-stage patent retrieval method considering the claim structure. In *ACM Transactions on Asian Language Information Processing (TALIP)* 4, 2, 190-206.
- [7] Moens M-F. (2005) *Automatic Indexing and Abstracting of Document Texts*. The Kluwer International Series on Information Retrieval Vol. 6. 343-347.
- [8] Sarasua L, Corremans G. (2000) Cross lingual issues in patent retrieval. In *Proceedings of the ACM SIGIR 2000 Workshop on Patent Retrieval*. ACM, New York.
- [9] Sheremetyeva S. (2003) Towards Designing Natural Language Interfaces. In *Proceedings of the 4th International Conference "Computational Linguistics and Intelligent Text Processing"* Mexico City, Mexico
- [10] Smeaton A, Sheridan P. (1991) Using morpho-syntactic language analysis in phrase matching. In *Proceedings of RIAO'91*. Barcelona, Spain, pp. 414-430.
- [11] <http://www.infonortics.com/chemical/ch04/slides/lambert-new.pdf>
- [12] <http://www.delphion.com>
- [13] <http://www.micropat.com>

APPENDICES

A. LIST OF GENERAL-PURPOSE TERMS IN THE ENGINEERING DOMAIN

N:system
N:apparatus
N:method
N:control
N:appliance
N:device
N:holder
N:tool
N:implement
N:member
N:body
N:means
N:assembly
N:frame

B. 20 MOST FREQUENT NOUNS IN THE CORPUS (GENERAL-PURPOSE TERMS ARE UNDERLINED)

4759 N:end
4089 N:device
3995 N:portion
3943 N:surface
3918 N:control
3448 N:position
3362 N:output
3160 N:system
3082 N:member
3036 N:apparatus
2996 N:current
2978 N:means
2919 N:frame
2737 N:material
2654 N:assembly
2650 N:air
2634 N:circuit
2566 N:support
2483 N:tool
2393 N:cylinder