Enriching Patent Search with External Keywords: a Feasibility Study

Ivelina Nikolova, Irina Temnikova, Galia Angelova
Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
iva@lml.bas.bg, irina.temnikova@gmail.com, galia@lml.bas.bg

Abstract

This article presents a feasibility study for retrieving Wikipedia articles matching patents' topics. The long term motivation behind it is to facilitate patent search by enriching patent indexing with relevant keywords found in external (terminological) resources, with their monolingual synonyms and multilingual translations. The similarity between patents and Wikipedia articles is measured using various filtering techniques and patent document sections. The most similar Wikipedia articles happen to be the closest ones to the respective patent in 33% of the cases, otherwise they are within the top 12 ranked articles.

1 Motivations and Related Work

Patent documents exhibit structure uniformity (Alberts et al., 2011) and have assigned classification codes but patents search is not a trivial task. This is due to the large number of patents available worldwide (forty millions) (Hunt et al., 2007) and the specific language genre. Usually the invention descriptions aim at covering the widest possible application area and are intentionally left very vague. Thus patents do not follow a pre-established terminology but rather are written according to the specific lexicon and style of each inventor (Alberts et al., 2011). Patent applications are published before the granting decision, therefore their titles and abstracts are intentionally left very general (Adams, 2010a). Moreover, the internationally used classification hierarchies vary among institutions and are periodically changed.

The present NLP technologies provide insufficient support to patent searchers' needs (Lupu et al., 2011; Adams, 2010a). Full-text search is the most preferred type of patent search while examining a patent application in order to establish its novelty, patentability, and infringement (Adams, 2010a). Search is done through iterative attempts, using synonyms in order to catch the alternative expressions each inventor may use to describe the same concept (Hunt et al., 2007). It is known that it can take up to 40 hours (in average 12) for a specialist to complete the search task for 15 queries in 100 documents, including a minimum of 5 minutes for a single query formulation (Joho et al., 2010). Another specific requirement is that patent searchers need the highest possible recall because a single relevant missed document can invalidate an otherwise sound patent (Lupu et al., 2011).

Our original idea is to use Wikipedia as a free, multilingual and constantly updated terminology resource, in order to enrich patent indexing with monolingual term synonyms and their translations in multiple languages. This would allow increasing patent search recall, and it is the solution we propose to recognizing vague and inventor-specific term definitions. Wikipedia is constantly updated; besides the multiple critiques to the reliability of Wikipedia articles[1], its peer-review nature repays for it (Giles, 2005). Thus the automatic recognition of relevant to the patent's topic Wikipedia articles is a first experimental step towards enriching patents indexing with Wikipedia terms. As many Wikipedia article titles are homonyms (usually described in disambiguation pages[2]), full-text article recognition is necessary.

Related Work in NLP for patents. Most of the NLP approaches contributing to patent search have been published in the CLEF-IP[3], TREC-CHEM[4] tracks, the NTCIR workshops patents tracks for Japanese, and in the PaIR[5] workshops. Lupu et al. (2011) provides a very good overview of the state-of-the-art of IR technologies for patents and how well they respond to the users' needs. Multilinguality in patents search is

[1] http://en.wikipedia.org/wiki/Reliability_of_Wikipedia
[2] http://en.wikipedia.org/wiki/Wave_%28disambiguation%29
[3] http://www.ifs.tuwien.ac.at/ clef-ip/index.html.
[4] http://www.ir-facility.org/trec-chem.
[5] http://www.ir-facility.org/pair-workshops.

Proceedings of Recent Advances in Natural Language Processing, pages 525–531,
Hissar, Bulgaria, 7-13 September 2013.

prevalently addressed by automatically translating whole patents in other languages into the language of the query. The existing approaches tackle a variety of specific patents retrieval tasks, ranging from patents language analysis (Shinmori, 2003), to patent retrieval evaluation (Lupu et al., 2011).

Among the closest to ours approaches is Pesenhofer et al. (2011), who assign new index terms to patents by retrieving relevant Wikipedia Science Portal pages. The difference with our work is that we plan to assign to patents as indexing terms only synonyms specified in the particular Wikipedia articles and translation equivalents from the linked pages, and that their approach takes into consideration only strictly scientific topics. Another relevant work is Magdy and Jones (2011) who generate synonyms for query terms using WordNet (Fellbaum, 1998). Compared to the approaches, currently known to us, the originality of our idea consists in the automatic generation of suggestions for (multilingual) synonyms with assigned similarity scores to be shown to the patentees when they perform patent searches.

2 Materials Used

The experimental dataset is a subset of patents from MAREC400k that belong to the MAREC corpus[6]. MAREC is a static collection of over 19 million patent documents provided as XML files, unifying 100,000 randomly selected patent applications and granted patents from four main patent authorities: the European Patent Office (EPO), the World Intellectual Property Organization, the US Patent and Trademark Office, and the Japan Patent Office. MAREC has been compiled specifically for NLP/IR/MT research by the IR Facility in Vienna [7]. We use only a subset of MAREC400k, which contains patents in English from the EPO collection, with the following subject categories (according to the International Patents Classification, IPC): *A43* – Footwear, *A44* – Haberdashery, Juwellery, *A45* – Hand or travelling articles, *A47* – Furniture and Domestic Articles, *G06* – Computing, *G07* – Checking Devices, and *G09* – Education, Cryptography.

We use only patent documents which contain the sections *Description* and *Claims* in addition to the patent *Invention title* and *Abstract*. A human judge collected our experimental corpus. He was asked to go manually through the patents, decide the topics and assign the most relevant Wikipedia articles to each of the patent documents as a whole, and to each of the patents' paragraphs, including claims. For this reason, in this experiment we use a restricted number of fifteen patent documents within the above-mentioned categories with length between 4 and 30 pages. It is known that most terms characterizing the invention are contained in the invention description and in the patent claims, while the patent title, abstract, and the context of the problem contain only very general information (2010a; 2010b).

In our experiment we used Wikipedia articles from the English Wikipedia. The corpus contains: *(i)* manually identified articles discussing the topics of the selected patents, with the best similarity match to the patents topics, 1-3 per patent (29); *(ii)* manually selected Wikipedia pages as distractors (articles discussing topic similar but not the same as the patent's ones), 2-20 per patent (153), and *(iii)* randomly selected Wikipedia articles (6,747).

All Wikipedia articles and patents are preprocessed within the GATE framework (Cunningham et al., 2011), (Cunningham et al., 2002), using the ANNIE processing resources (Cunningham et al., 2002). The XML- and MediaWiki-markups are ignored, the text is lemmatized and the calculation of similarity is done on lemmas only. Stopwords were marked and later we made experiments with both corpora - documents containing stopwords and documents with removed stopwords.

3 Experiments Design and Results

Our study includes identifying the closest Wikipedia article match with the currently processed patent document. As often there are Wikipedia pages with homonym titles, the "closest" article cannot be identified only by its title, e.g. *seat* (where one sits) vs *SEAT* (a car brand). In order to overcome homonym titles disambiguation, our approach uses the whole text of the Wikipedia articles, the patent document description, the patent categories and the patent claims. After calculating the similarity between a number of patent texts (or patent parts) and Wikipedia pages, we check if the closest match automatically identified by our method, corresponds to the closest match previously identified by a human judge.

[6] http://www.ir-facility.org/prototypes/marec
[7] http://www.ir-facility.org/home

Experiment 1 – Patent descriptions vs Wikipedia articles. To determine the best method for text similarity calculation, we performed several experiments with different algorithms for calculating the semantic distance between patent description texts and Wikipedia articles. We used the DKPro Similarity Framework (Bär et al., 2012) and applied most of the similarity measures available there, among which are *WordNGramJaccard measure*, *ExactStringMatch comparator*, *JaroSecondString comparator*, *JaroWinklerSecondString comparator*, *LevenshteinSecondString comparator*, and *LongestCommonSubstring comparator*. The best results were obtained with the classical *CosineSimilarity measure*. We used it in the study presented here.

A round of experiments has been done without using a stopwords list. The similarity was calculated on the basis of the words' lemmas. Although the highest similarity scores were quite close to 1, the results were rather discouraging, the documents having high scores were often not similar to the patents at all and the manually assigned as "most similar" documents were not given a high score. This is why we decided to use a stop word list in the further experiments, and it proved to be a better choice.

We made 2 separate runs of the similarity calculations. In Run 1 we measured the semantic distance between patent descriptions and a number of manually selected Wikipedia articles, annotated with the boolean values - *similar* or *distractor*. The results are illustrated in Table 1, Run 1, for each patent description. The 2nd column shows the position where the manually pre-defined "most similar" pages appeared among the top 20 highly ranked relevant Wikipedia pages. In all cases, except for the 11th patent, the "most similar" pages are recognised, and in the 3rd, 7th, 8th and 13th case they have the highest ranking. In the 4th and 12th cases, one of the Wikipedia articles was given highest score and in the rest of the cases the correct articles were with lower rank but still within the top 20 results. Unfortunately we see that the Wikipedia pages, intentionally selected as distractors, appear as highly similar documents as well, which means that the mere computation of similarity using bag-of-words techniques at this level is insufficient to ensure proper disambiguation.

In Run 2 we added some 6,747 randomly selected Wikipedia pages to the set of manually annotated pages. Many of these documents were given pretty high similarity score although they were irrelevant. Often they were about people, geographic locations and landmarks which are irrelevant to the patent data. We decided to remove such pages before running the similarity algorithm. We filtered them by their Wikipedia category and we ended up with 1,465 randomly selected Wikipedia pages. We note that the Wikipedia category tree is not consistently developed and it is not trivial to select all categories matching these types of articles thus some might be omitted. By augmenting the set of Wikipedia articles our goal was to check whether the algorithms will perform consistently and will assign higher score to the same pages as it did in the first run. The results are shown on Table 1, Run 2. We see that for the 2nd, 3rd, 4th, 7th, 8th, 12th and 13th patent the results are the same as in the case of the manually selected Wikipedia pages. For some cases there are slight shifts in the ranking, and for patent #11, the "most similar" pages do not appear at all among the top 20 closest documents in both Run 1 and Run 2. As a reason for that we see that the patent text is rather a functional description of the entertainment machine and the closest Wikipedia articles explain about the history and application of the entertainment machines.

The upper part of Table 2 shows the similarity scores calculated for the patent EP-0073116-A2 *Integrated data processing circuits* and the Wikipedia pages. The full patent text can be seen at the EPO site. The manually selected matching pages from Wikipedia are in bold. They appear in the top ranked results but without significantly higher similarity score. In addition to the manually selected pages here *Asynchronous circuit* and *Intel MCS-51* appear with very high similarity score. Indeed they are similar to the topic because *Intel MSC-51* is an implementation of integrated circuit and asynchronous circuit is also type of integrated circuit - sequential digital logic circuit. This is an example of gathering new potential indexing keyterms. The articles *Clock* and *Multiplication* have also been given pretty high score. Although the expert did not select them as closest matches to this patent, he did select them as "most similar" to some of the patent paragraphs, which means that they are also true positives and are appropriate to describe this patent.

The lower part of Table 2 shows the similarity

Pat. id	Rank in top 20 res	Wiki docs	Rank in top 20 res	Wiki docs	Rank in top 20 res	Wiki docs	Rank in top 20 res	Wiki docs	Rank in top 20 res	Wiki docs	Rank in top 20 res	Wiki docs	Rank in top 20 res
	Run 1		Run 2		Run 3		Run 4		Run 5		Run 6		Run 7
1.	2	156	3	1 465	N/A	1 465	3	1 465	3	1 465	3	1 465	2
2.	2, 4	156	2, 4	1 465	1, 3	1 465	2, 4	1 465	2, 8	1 465	2	1 465	2, 4
3.	1, 2	156	1, 2	1 465	1, 2	1 465	1, 2	1 465	1, 2	1 465	1, 2	1 465	1, 2
4.	1, 10	156	1, 10	1 465	2, 20	1 465	1, 10	1 465	1, 12	1 465	1	1 465	1, 12
5.	3, 12, 15	156	3	1 465	3, 13	1 465	3	1 465	2, 15	1 465	2	1 465	2, 12
6.	1, 2	156	1	1 465	3	1 465	1	1 465	2	1 465	2	1 465	1, 11
7.	1, 2	156	1, 2	1 465	1, 2	1 465	11, 20	1 465	1, 2	1 465	1, 2	1 465	1, 2
8.	1	156	1	1 465	6	1 465	1	1 465	1	1 465	1	1 465	1
9.	7	156	7	1 465	10	1 465	7	1 465	8	1 465	8	1 465	1
10.	1, 6	156	1, 10	1 465	6, 7	1 465	1, 10	1 465	1, 8	1 465	1, 8	1 465	1, 7
11.	N/A	156	N/A	1 465	7	1 465	N/A	1 465	N/A	1 465	N/A	1 465	N/A
12.	1, 4	156	1, 4	1 465	17	1 465	1, 4	1 465	1, 4	1 465	1, 4	1 465	1, 4
13.	1	156	1	1 465	6	1 465	1	1 465	1	1 465	1	1 465	1
14.	4, 6	156	7, 9	1 465	N/A	1 465	N/A	1 465	8, 10	1 465	8, 10	1 465	6, 8
15.	7, 11, 16	156	7, 11	1 465	6, 11	1 465	7, 11	1 465	6, 10	1 465	7, 11	1 465	6, 11

Table 1: Rank of the most similar documents according to cosine measure.
Run 1 - Patent descriptions and manually selected Wiki-articles; Run 2 - patent descriptions and both, manually and randomly selected Wiki-articles; Run 3 - patent categories and both, manually and randomly selected Wiki-articles; Run 4 - patent claims and both, manually and randomly selected Wiki-articles; Run 5 - combined patent description with claims and both, manually and randomly selected Wiki-articles; Run 6 - combined patent categories, description, claims and both, manually and randomly selected Wiki-articles; Run 7 - weighted similarity between Wiki-articles and a patent considering the scores from Runs 2–6.

Invention title: Integrated data processing circuits. Patent ID: EP-0073116-A2, Category: G06F.	
Wikipedia match: **Integrated circuit; Very-large-scale integration.**	
Run 1	Run 2
Rank of Wiki-pages sorted by cosine similarity:	
1. Asynchronous circuit (0.218) 2. Clock (0.207) 3. Multiplication (0.170) **4. Integrated circuit** (0.151) 5. Computer (0.151) **6. Very-large-scale integration** (0.147) ...	1. Intel MCS-51 (0.246) 2. Glia limitans (0.231) 3. Pennales (0.224) 4. Asynchronous circuit (0.218) 5. Clock (0.207) 6. Multiplication (0.170) **7. Integrated circuit** (0.151) 8. Computer (0.1506) **9. Very-large-scale integration** (0.147) ...

Invention title: Folding table or like structure. Patent ID: EP-0105957-A1, Category: A47B.	
Wikipedia match: **Table (furniture); Folding table.**	
Run 1	Run 2
Rank of Wiki-pages sorted by cosine similarity:	
1. Table (furniture) (0.509) **2. Folding table** (0.489) 3. Table (database) (0.339) 4. Table (parliamentary procedure) (0.270)	**1. Table (furniture)** (0.509) **2. Folding table** (0.489) 3. Table (database) (0.339) 4. Table (parliamentary 4. procedure) (0.270)

Table 2: Run 1 and 2 with patents EP-0073116-A2 and EP-0105957-A1

scores calculated for patent EP-0105957-A1. In this case the matching Wikipedia pages are the top closest results. We can view their Wikipedia categories as potential indices of EP-0105957-A1 as well: for the article *Table (furniture)* in Wikipedia these are *Tables (furniture)* and *Furniture*. So the latter term can be shown to a patent searcher as a potential descriptor. It reveals the semantics of EP-0105957-A1 despite the fact that it does not appear in the patent text at all.

Experiment 2 – Patent categories vs Wikipedia articles. We decided to observe the similarity between patents and Wikipedia articles from one more perspective: document categories versus document text contents. We extracted all categories of each patent (varying between 1 to 15 per patent), transformed their reference numbers into the titles of the categories, and pre-processed them as a regular text document. Then we measured the similarity between these lemmatized texts and the Wikipedia articles.

We decided to use patents categories and Wikipedia texts, rather than the opposite (patent description and Wikipedia categories), because the IPO categorical tree is precisely elicited and the categories which are assigned to each patent are carefully chosen to make the patent easy to retrieve during search. Whereas Wikipedia categories are not really strictly organised and the depth of the categorical tree varies a lot from branch to branch. In Wikipedia it is very common that some articles on a topic, which is not popular, have only few categories listed (one or two), even if there are many other appropriate ones existing. In the same time articles like Barack Obama have 50 assigned categories. The process of assigning categories to patents is somehow better regulated.

We tested also this approach with and without using stop words and again only when we removed the stop words we could obtain meaningful results. The categories files are rather short, containing essential information and removing the stop words emphasises even more the keywords they contain. The presented results are only from the experiment when stop words are removed. We measured the similarity between the patent categories and the whole set of Wikipedia articles including manually selected and randomly added ones. The results are shown on Table 1, run 3.

Experiment 3 – Patent claims vs Wikipedia articles. We took also a third perspective in measuring the similarity between these two types of documents. We extracted all patent claims (varying between 4 and 36 per document) and preprocessed them as regular text documents. These differ from the patent description as they contain the synthesized essence of the invention, in bullet points, while the patents description contains also an overview of the problem background, and it is thus much more general. We applied the same similarity measure between these lemmatized texts and the whole collection of Wikipedia articles including the manually and randomly selected ones. The results, obtained after stop words removal, are presented on Table 1, Run 4.

Experiment 4 – Comparison of patent subsections with full text Wikipedia articles. The aim is to *(i)* find better matches to specific document sub-parts, describing specific techniques or methods, which may be used in other inventions. And thus adding new keywords describing these sub-parts, we augment the chance that the patent searcher will find those in order to prevent any infringement of the rights of previous patents. On the other hand, *(ii)* test if this helps to improve the match of the whole document to Wikipedia articles. Our hypothesis is that the description paragraphs would have more diverse matches than the claims, as manual analysis has shown that each claim tends to be more precise and mentions several times the object of the invention. Thus, by splitting the descriptions into paragraphs, we expect to find more Wikipedia article matches to the same patent. Further, a human judge has manually identified the best matches for some claims or paragraphs, to test if the short text of a paragraph is enough to have similarity between it and the appropriate Wikipedia article.

The motivation for this approach is that it often happens, that the same invention has parts describing specific and very concrete technologies, borrowed from other fields. For example, a patent application, describing a technology improving integrated data processing circuits (patent reference number EP-0073116-A2), can contain paragraphs, discussion specifically multiplication specificities, and clocks, operating with phase difference. While a patent application, discussing a spring seat invention (patent reference number EP-0090622-A1), can include paragraphs, discussing the interactions between human's ischial tuberosities with seating surfaces, or using webbing and clamps in a specific way to keep together parts of the invention. As, sometimes, the claims of a patent may contain these specific technologies, as part of the invention, it is necessary to check if they have not been used in previously granted patent applications or, if used, whether mentioning them in the claim can infringe previous patents rights. We consider that retrieving more patents, discussing these topics, will assist patent specialists in reviewing all possible applications which are related to this invention.

Initially we set-up a paragraph to be any sequence of characters between two new lines. These turned to be often very short, sometimes section titles and in general not informative enough to have a meaningful comparison with a full Wikipedia article. The results were rather discouraging and then we set up a minimum paragraph size of 500 chars. Thus paragraphs which were shorter than 500 chars were added to the next paragraph. In this Run again removing the stop words gave better results.

The obtained similarity scores between patent paragraphs and Wikipedia articles resemble quite a lot to the results obtained from the full patent descriptions and Wikipedia articles. The manually selected "most similar" Wikipedia articles are ranked within the top 20 results, however it is hard to distinguish them from the distractor articles. Indeed some Wikipedia articles which are similar with concrete paragraphs receive higher similarity score in this experiment, but it turns that they receive high similarity score also with the whole patent description. Some of the results are shown on Table 3. The expert has selected *Multiplication* and *Clock* as "most similar" pages to the 3rd paragraph of the patent however the rest of the top ranked articles are also true positives.

Experiment 5 – Combined patent parts vs Wikipedia articles. After running all comparisons of the separate patent parts we observed the results and decided to combine these parts and compare them once again to all Wikipedia articles. We made two separate runs. Once we combined only the patent description and claims because we noticed that the results when using these

Invention title: Integrated data processing circuits. Patent ID: EP-0073116-A2, Category: G06F Paragraph: 3
Wikipedia match: **Integrated circuit**, **Very-large-scale integration**
Rank of the Wikipedia pages sorted by cosine similarity: 1. **Multiplication** (0.169) 2. Asynchronous circuit (0.156) 3. Integrated circuit (0.139) 4. Very-large-scale integration (0.123) 5. **Clock** (0.122)

Table 3: Top results from matching paragraph 3 of EP-0073116-A2 with all Wikipedia articles.

two parts (Run 1,2 and 3) are more consistent that the ones obtained by the patent categories (Run 4). Then we combined also patent categories, description and claims (Run 6). We observed the change in the similarity score and ranking between the patents and the manually selected Wikipedia matches. The results from this experiment are presented on Table 1, Run 5 and Run 6.

Experiment 6 – Weighted Scoring of Wikipedia articles. To filter out the results obtained from all these experiments we calculated the weight of each Wikipedia article according to each patent, using the score obtained by the similarity algorithm and the number of times a Wikipedia article is ranked among the top 20 ones:

$$Weight = \sum_{i=1}^{i=n} Rank_i * Score_i$$

where n is the number of experiments, i.e. n=5 excluding Run 1 (without stopwords).

This way we give preference to the articles which appear more often than the others in the top results and to the ones with higher score. Although this technique is rather simple it allowed us to restrict the true positives within the top 12 results. In 8 of the cases they were within the top 5 results. We would like to mention that the fact that some manually selected Wikipedia articles appear with lower rank, often means that there are other very similar articles which were not selected by the human judge as such, and they appear with higher rank, and they are also appropriate to be used for indexing of that patent.

4 Discussion and conclusion

The results on Table 1 show the change in the ranking of the ”most similar” manually selected Wikipedia articles when calculating similarity between different parts of the patents and Wikipedia articles. In Run 1 only 156 Wikipedia articles are used, in Run 2 - 10 times more (1 465), and there are still only slight differences in the ranking of the ”most similar” articles in both runs. This stability in the performance of the cosine similarity algorithm in this task is encouraging for applying it for even bigger data sets. We see that the Wikipedia articles, which receive high similarity score and rank to some patent, retain it in all experiments (with claims, description, combined). The only experiment which gives somehow inconsistent results is (Run 3) where we map patent categories to Wikipedia articles. This must be due to the fact that patent categories are short expressions with rather general wording. Thus our feasibility study shows that the identification of the closest match is possible, but it is difficult to distinguish between closest and close results. In general the results are promising since the recall in patent search is more important than the precision, and thus the noise is not so disturbing. However much work remains to be done for improving the computation paradigm and refining the precision. Further, we aim at extracting synonyms and translation equivalents to enrich patents indexing, and this requires additional experiments with real users.

Another challenge is to elaborate the initial filtering of the Wikipedia articles in order to better restrict the categories of Wikipedia pages. For instance, pages for cities, states and provinces contain long descriptions about industries, communications etc. and therefore they might be identified as ”similar” to various patents, so it is be reasonable to remove such pages from the experiment at all. Future work includes also experiments with assigning weight to the words in the patent description and claims, and processing multiword expressions. Last but not least, employment of multilinguality in decision making regarding similarity is possible as well. Wikipedia is multilingual and patents contain titles and abstracts in several languages, so patent fragments in another language can be used to calculate similarity with Wikipedia pages in the corresponding language.

Acknowledgements. The research work presented in this article is partially supported by the grant 316087 AComIn ”Advanced Computing for Innovation”. It is also related to the COST Action IC 1002 MUMIA ”Multilingual and Multifaceted Interactive Information Access”.

References

Adams Stephen 2010. *The text, the full text and nothing but the text: Part 1Standards for creating textual information in patent documents and general search implications*. World Pat Inf 32:22 29.

Adams Stephen. 2010. *The text, the full text and nothing but the text: Part 2The main specification, searching challenges and survey of availability.*. World Pat Inf 32:120128.

Alberts, D., C. B.Yang, D. Fobare-DePonio, K. Koubek, S. Robins, M. Rodgers, E. Simmons, and D. DeMarco. 2011. *Current Challenges in Patent Information Retrieval, Chapter 1: Introduction to Patent Searching-Practical Experience and Requirements for Searching the Patent Space.* The Information Retrieval Series, Volume 29. Springer-Verlag Berlin Heidelberg.

Bär, Daniel, Chris Biemann, Iryna Gurevych and Torsten Zesch. *UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures* In Proceedings of the 6th Int'l Workshop on Semantic Evaluation, in conjunction with the 1st Joint Conf. on Lexical and Computational Semantics pages 435-440, June 2012, Montreal, Canada.

Chen, L., Tokuda, N., & Adachi, H. 2003. *A patent document retrieval system addressing both semantic and syntactic properties*. In Proceedings of the ACL-2003 workshop on Patent corpus processing-Volume 20 (pp. 1-6). Association for Computational Linguistics..

Choi, Sung-Kwon, Oh-Woog Kwon, Ki-Young Lee, Yoon-Hyung Roh, and Young-Gil Kim. 2007. *Customizing an English-Korean Machine Translation System for Patent Translation*. In The 21st Pacific Asia Conference on Language, Information and Computation (PACLIC 21), pp. 105-114.

Cunningham, Hamish and Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li and Wim Peters. 2011. *Text Processing with GATE (Version 6)*. ISBN 978-0956599315, 2011, http://tinyurl.com/gatebook

Cunningham, Hamish, Diana Maynard, Kalina Bontcheva and Valentin Tablan. *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*, In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), 2002

H. Cunningham, Valentin Tablan, A. Roberts, K. Bontcheva (2013) *Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics.* PLoS Comput Biol 9(2): e1002854. doi:10.1371/journal.pcbi.1002854 http://tinyurl.com/gate-life-sci/

Fellbaum, C. 1998. *WordNet: an Electronic Lexical Database*. MIT Press.

Joho, Hideo, Leif A. Azzopardi, and Wim Vanderbauwhede. 2010. *A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements*. Proceedings of the third symposium on Information interaction in context, pp. 13-24. ACM.

Hunt, David, Long Nguyen, and Matthew Rodgers. (Eds). 2007. *Patent searching: Tools & techniques*. Wiley.

Giles, Jim. 2005. *Internet encyclopaedias go head to head*. Nature, vol. 438, 7070, pp.900-901 Nature Publishing Group.

Lu, Bin, and Benjamin K. Tsou 2009. *Towards bilingual term extraction in comparable patents*. In Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC), pp. 755-762.

Lupu, Mihai, Katja Mayer, John Tait, and Anthony Trippe. (Eds). 2011. *Current Challenges in Patent Information Retrieval*. The Information Retrieval Series, Volume 29. Springer-Verlag Berlin Heidelberg.

Magdy, W., & Jones, G. J. 2011. *A study on query expansion methods for patent retrieval*. In Proceedings of the 4th workshop on Patent information retrieval (pp. 19-24). ACM.

Andreas Pesenhofer, Helmut Berger, and Michael Dittenbach. 2011. *Current Challenges in Patent Information Retrieval, Chapter 18: IOffering New Insights by Harmonizing Patents, Taxonomies and Linked Data*. The Information Retrieval Series, Volume 29. Springer-Verlag Berlin Heidelberg.

Sheremetyeva, Svetlana, Sergei Nirenburg, and Irene Nirenburg. 1996. *Generating patent claims from interactive input*. In Proceedings of the Eighth International Workshop on Natural Language Generation, pp. 61-70.

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-Volume 20 (pp. 56-65). Association for Computational Linguistics.

Wood, Andrew, Kate Struthers, and Uk Edinburgh 2010. *Pathology education, Wikipedia and the Net generation.*. Medicine 38 (2010): 868-878.