An Investigation of Decompounding for Cross-Language Patent Search

Johannes Leveling, Walid Magdy and Gareth J. F. Jones
School of Computing, CNGL
Dublin City University
Dublin, Ireland
{jleveling, wmagdy, gjones}@computing.dcu.ie

ABSTRACT

Decompounding has been found to improve information retrieval (IR) effectiveness in general domains for languages such as German or Dutch. We investigate if cross-language patent retrieval can profit from decompounding. This poses two challenges: i) There may be few resources such as parallel corpora available for training an machine translation system for a compounding language. ii) Patents have a specific writing style and vocabulary ("patentese"), which may affect the performance of decompounding and translation methods. Experiments on data from the CLEF-IP 2010 task show that decompounding patents for translation can overcome out-of-vocabulary problems (OOV) and that decompounding improves IR performance significantly for small training corpora.

Categories and Subject Descriptors

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval—*Query formulation*; H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing—*Linguistic processing*

General Terms

Experimentation, Performance, Measurement

Keywords

Patent Retrieval, Decompounding

1. INTRODUCTION

Compounding languages such as German or Dutch allow combining simple words into complex words by concatenating them. In contrast to English, these compounds are written as single words. Splitting compounds into their constituent parts (i.e. decompounding) has been found to improve IR effectiveness, because it can overcome vocabulary mismatches [1, 2, 6].

An area of increasing interest in IR is prior art patent search, which is concerned with finding all relevant patents for a patent application. Since patents are often written in different languages, cross-language information retrieval (CLIR) is usually an essential component of effective patent search. For patent search in compounding languages, the CLIR effectiveness is usually lower than for other language pairs [3, 7]. This can be attributed to the presence of compounds, which leads to higher rates of OOV compound terms. OOV terms cannot be translated, which results in missing some portion of the query text and degrades IR effectiveness.

In this paper we apply decompounding on German patent topics in the cross-language patent search task from the CLEF-IP 2010 track and investigate machine translation (MT) quality by examining retrieval performance. The results show that decompounding improves retrieval performance significantly for small training corpora, or for corpora with high OOV rates.

2. THE DECOMPOUNDING APPROACH

There has been little research on decompounding for patent search and for training MT systems. Koehn and Knight [4] train decompounding for MT using knowledge from parallel corpora, preventing incorrect decompounding when there is a one-to-one correspondence between two words in different languages. Jochim et al. [3] apply dictionary-based translation for cross-language patent search and expand monolingual queries with their translations. They conclude that translation could help patent retrieval, but not always.

In this paper, decompounding German words is realized by an approach which has been employed in domain-specific CLIR [2]. The decompounding is based on selecting the decomposition with the smallest number of words and the highest decomposition probability. A decomposition probability is defined as the product of constituent probabilities, which are estimated by the collection frequency of a word divided by the number of all words in a training collection. The training collection contains the English 3M sentence corpus from the Leipzig corpora list[1] and a random sample of 800k sentences from German patents in the CLEF-IP collection.

We evaluated the decompounding based on a gold standard corpus (GSC) of 2000 random sentences extracted from German patents. The GSC was manually annotated with the correct decomposition of words. It contains 27,932 unique words and 318k words in total. We found that spelling errors in the patent texts occur frequently, possibly resulting from the OCR source of documents. Spelling errors have also been manually decompounded in the GSC. In addition, 12.7% of the word forms in the annotated corpus are chemical formulas or substance names, which indicates the domain-specific nature of patents. In the GSC, chemical formulas are decompounded only when the head noun is a German word. For example, *Methylrest* (methyl radical) has the head noun *Rest* (radical).

This decompounding method achieves 95.0% accuracy (the percentage of correctly decompounded words) measured over all words in the annotated GSC and 81.4% accuracy for unique words. Decompounding the GSC increases the total number of words by 16.3%, while the number of hapax legomena (words occurring only once)

[1] http://corpora.uni-leipzig.de/

Table 1: Patent retrieval results for corpora.

	No decompounding		Decompounding	
MT Corpus Size	PRES	OOV	PRES	OOV
500K	0.486	10.2%	0.476	1.1%
50K	0.444	20.9%	0.479	3.7%
5K	0.360	40.8%	0.450	12.6%

is reduced by 48.8%, compared to the original GSC. This illustrates that compounding is a productive process in German.

3. EXPERIMENTS AND RESULTS

The cross-language search task in CLEF-IP 2010 is adapted for our patent retrieval experiments [7]. The main objective is to find patents in an English collection that are relevant to patent applications filed in German. The collection consists of 1.35M patents from the European Patent Office with 69% of them exclusively in English and 31% in German and French. The German and French patents in the collections often have sections manually translated into English, including the patent title, abstract, and claims. For our experiments, all the English text of the collection was indexed to create a monolingual index.

For the translation process, the MaTrEx MT system[2] was used to translate the 89 German patent topics from the small topic set (300 topics in English, French, and German) provided by the CLEF-IP 2010. A random set of parallel sentences in English and German from the title and claims sections of patents was extracted to train MaTrEx. Different sizes of a training corpus, namely 500k, 50k, and 5k parallel sentences, were used to investigate the effect of a training corpus on the translation quality which can be indirectly observed by retrieval effectiveness. For comparison, the German sentences in the three training corpora were decompounded to create another three training corpora, which creates a total of 6 translation models for the MT system. For the German topics, decompounding was applied too to create a decompounded version of the topics to be translated with the decompounded translation models.

Translated patent topics were processed to form queries by adding terms occurring more than twice in the title, abstract, description, and claims sections combined and all bigrams that occur more than three times, using the term frequency as a weight for these terms [7]. The INDRI[3] toolkit was used to index and search the patents.

Table 1 shows PRES results [5] and OOV for cross-language patent search for different MT training corpus sizes, using the 89 German patent topics. PRES scores increase when larger MT corpora are used which are not decompounded. However, significance tests (t-test, $p < 0.5$) show that the PRES scores using the decompounded model for 500k or 5k are not significantly different and that results for the 5k and 50k model are significantly better for the respective decompounded version (see Table 1).

Unexpectedly, results are significantly indistinguishable for the decompounded and the original 500k training corpus. To find an explanation, we investigated several topics for which PRES has decreased using MT with decompounding. We illustrate our findings on topic PAC-199, for which PRES decreases from 0.843 to 0.691 (mean average precision decreases from 0.414 to 0.158). The word *schwerbrennbar* (flame retardant or flame resistant) is incorrectly translated as heavy combustible, which results in matching non-relevant documents, compared to no additional non-relevant matches when the word is not translated at all. More importantly, the chemical formula *Methylvinylsiloxan* (term frequency 35 vs. 0) is split into the constituents *Methyl* (20 vs. 103), *Vinyl* (15 vs. 118), and *Siloxan* (7 vs. 43). The numbers in brackets show the changes in term frequency for the unprocessed topic versus the topic after decompounding (topics correspond to full patents). This indicates that splitting some compounds, especially highly frequent chemical formulas in patents, can result in performance loss. The high PRES scores for the 5k and 50k decompounded models can be at least partially explained by the lower OOV rate in the decompounded version (Table 1). For the 500k training model, the OOV rate decreases, but the positive effect of decompounding might be outweighed by over-splitting of some compounds.

4. CONCLUSION

When using smaller MT training corpora or corpora with high OOV rates, decompounding shows higher performance compared to not decompounding. This effect is important for training MT systems for specific domains where only small parallel corpora are available for training MT systems, i.e. less training data is needed for good CLIR performance when decompounding is used to overcome OOV problems, and for training MT systems for languages with few linguistic resources.

The effect of incorrectly translating a constituent word of a compound is similar to a topic drift in blind relevance feedback, when adding query terms may result in a loss of precision. In patent search, where queries are formed by patent documents, incorrectly decompounded or incorrectly translated constituent words may result in a much higher query term frequency (and thus, a higher term weight in the query), causing a loss in retrieval effectiveness. This result is in line with observations made by Koehn and Knight [4], who conclude that "eager splitting fares abysmally".

Future work includes investigating methods to identify technical terms and chemical formulas in patents to treat them differently for decompounding and/or translation.

Acknowledgments

This research is supported by the Science Foundation of Ireland (grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (`http://www.cngl.ie/`).

5. REFERENCES

[1] M. Braschler and B. Ripplinger. How effective is stemming and decompounding for German text retrieval? *Inf. Retr.*, 7(3-4):291–316, 2004.

[2] A. Chen and F. C. Gey. Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Inf. Retr.*, 7(1–2):149–182, 2004.

[3] C. Jochim, H. Lioma, H. Schütze, S. Koch, and T. Ertl. Preliminary study into query translation for patent retrieval. In *PaIR '10*, Toronto, Canada, 2010.

[4] P. Koehn and K. Knight. Empirical methods for compound splitting. In *EACL '03*, pages 187–193, Stroudsburg, PA, USA, 2003. ACL.

[5] W. Magdy and G. J. F. Jones. PRES: a score metric for evaluating recall-oriented information retrieval applications. In *SIGIR 2010*, Geneva, Switzerland, 2010.

[6] C. Monz and M. de Rijke. Shallow morphological analysis in monolingual information retrieval for Dutch, German, and Italian. In *CLEF '01*, pages 262–277. Springer, 2002.

[7] F. Piroi. CLEF-IP 2010: Retrieval experiments in the intellectual property domain. In *CLEF 2010*, Padua, Italy, 2010.

[2] `http://www.openmatrex.org/`

[3] `http://www.lemurproject.org/indri/`