Enhanced Thesaurus Terms Extraction for Document Indexing

Frane Šarić, Jan Šnajder, Bojana Dalbelo Bašić, Hrvoje Eklić
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Croatia
E-mail:{Frane.Saric, Jan.Snajder, Bojana.Dalbelo, Hrvoje.Eklic}@fer.hr

Abstract. *In this paper we present an enhanced method for the thesaurus term extraction regarded as the main support to a semi-automatic indexing system. The enhancement is achieved by neutralising the effect of language morphology applying lemmatisation on both the text and the thesaurus, and by implementing an efficient recursive algorithm for term extraction. Formal definition and statistical evaluation of the experimental results of the proposed method for thesaurus term extraction are given. The need for disambiguation methods and the effect of lemmatisation in the realm of thesaurus term extraction are discussed.*

Keywords. Information retrieval, term extraction, NLP, lemmatisation, Eurovoc.

1. Introduction

Finding documents on the Web or in large document databases that are relevant for user's queries is the primary research topic in the field of information retrieval (IR). Document indexing is the process of assigning one or more key phrases that describe the content of the document in order to facilitate IR. These key phrases (called *terms* or *descriptors*) usually belong to a finite set of phrases arranged in the form of a controlled vocabulary or thesaurus. Thesauri contain additional information about term relationships, such as: related terms, hypernyms, hyponyms, etc., thus providing the means to control recall and precision of searches [8]. Examples of widely used thesauri are the Eurovoc (EUROpean VOCabulary) [7] and NASA Thesaurus [10].

Manual indexing is a time consuming, expensive intellectual task and is often inhomogeneous due to diverse background knowledge and expertise of human indexers. The task of building semi-automatic and automatic systems, which aim to decrease the burden of work borne by indexers, has recently attracted interest in the research community [4], [13], [14]. Automatic indexing systems still do not achieve the performance of human indexers, so semi-automatic systems are widely used (CINDEX, MACREX, MAI [10]).

In this paper we present a method for thesaurus term extraction regarded as the main support to semi-automatic indexing system. Term extraction is a process of finding all verbatim occurrences of all terms in the text. Our method of term extraction is a part of CADIS [9], and is meant to facilitate finding those terms that are explicitly contained in a document, although the document does not necessarily need to be indexed with the extracted terms.

The process of term extraction gives rise to some practical problems concerning term variation, such as morphological, lexical, structural, etc. An example of a method which deals with term variation for English is presented in [11]. We restrict our work to variation due to inflectional morphology, which makes the words appear in various forms. We can implicitly cope with some other types of term variation by using thesauri which encode relationships between synonyms.

Recall of term extraction suffers in documents written in morphologically rich languages (such as Croatian), so the effects of morphology have to be neutralised. The method described in the paper enhances the process of term extraction in two aspects. It efficiently tackles the problem of language

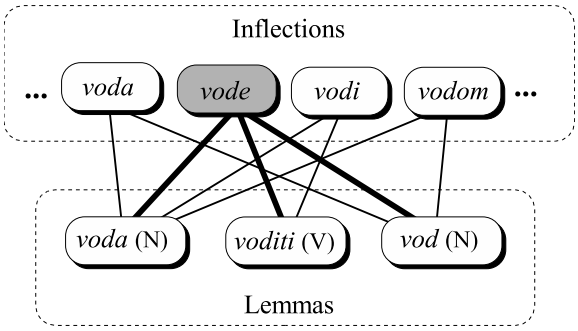

Figure 1: Croatian word *vode* has three lemmas: *voda* (water), *vod* (a duct or a squad) and *voditi* (to conduct, to lead).

morphology by applying lemmatisation, the most prominent natural language processing (NLP) technique used for indexing, on both the text and the thesaurus. The process of extraction is further enhanced by identifying and ignoring some cases in which terms are considered irrelevant.

The paper is structured as follows. In Section 2, we address the problem of language morphology, and describe our approach to lemmatisation. In Section 3, formal definition of our method is given. Finally, in Section 4 we present the statistical evaluation of experimental results of Eurovoc term extraction on a set of parallel documents written in Croatian and English.

2. Lemmatisation

2.1. The problem of morphology

When extracting terms from text documents, the effects of language morphology have to be taken into account. Relevant to the task of term extraction are the effects of inflectional morphology. It describes how from basic word form (the *lemma*) different word forms are generated in order to express grammatical features (e.g. number, case, gender, degree etc., depending on the word's part-of-speech). If term extraction were performed by literal string matching, various inflections of a term would not be found in the document, resulting in decreased recall.

To neutralise the effect of inflective morphology, each word form has to be *lemmatised*, i.e. a lemma for a given inflected form has to be found. Lemmatisation procedures range from purely algorithmic (rule-driven) to lexicon-based (relying on queries made to a morphological lexicon). For highly inflected languages the latter approach is more common. The *morphological lexicon* typically relates all inflected forms of a word to its lemma. The construction of a morphological lexicon is a labour intensive task. To facilitate the process, various automatic and semi-automatic procedures based on lexical acquisition from corpora have been developed [3], [6], [12], [15].

In our work, contrary to the usual practice, the process of lemmatisation does not imply word disambiguation. Instead, lemmatisation of an ambiguous word results in more than one lemma. Ambiguity considered here is called *homography* – the case when two or more lemmas have overlapping forms. A notorious example in English is the word *saw*, which can be a noun (a tool used for cutting) or the past tense of the verb *see*. An example in Croatian is the word *vode* as a feminine noun *voda* (water), a masculine noun *vod* (a duct or a squad), or a verb *voditi* (to conduct, to lead), as shown in Fig. 1.

2.2. Our approach to lemmatisation

In our work lemmatisation of English and Croatian documents and terms is performed using appropriate morphological lexicons.

For lemmatisation of English, one of many publicly available lexicons was used [1]. It contains over $250,000$ forms assigned to more than $100,000$ lemmas, with 4.8% ambiguous forms due to homography.

For lemmatising Croatian, a morphological lexicon constructed by a rule-based automatic acquisition [15] from a subsection of Croatian National Corpus [5] totaling 10^7 words was used. The obtained lexicon contains over $500,000$ forms assigned to more than $30,000$ lemmas. Degree of homography in lexicon is 5.1%.

In the process of term extraction, both precision and recall depend on the linguistic validity and the coverage of the morphological

lexicon used for lemmatisation. Estimates for lexicons used in our experiments will be given in Section 4.

3. Term Extraction

3.1. Formal definition

In order to define term extraction formally, word and term matching need to be defined first. Let W be a set of all words and $L : W \rightarrow \wp(W)$ denote a function that maps each word to the set of its lemmas, e.g. $L(vode) = \{vod, voda, voditi\}$. If we are not using lemmatisation or word w is not listed in the lexicon (which is usually the case for non-inflective words), then $L(w) = \{w\}$. We define words w_1 and w_2 to *match* iff $L(w_1) \cap L(w_2) \neq \emptyset$, i.e. both words are inflections of a common lemma.

We represent a term as a list of words $(t_1, \ldots, t_m)$, and a portion of text with no intervening punctuation as list of words $(w_1, \ldots, w_n)$. We define three relations that are relevant for term extraction. Term $(t_1, \ldots, t_m)$ *matches* a list of words $(w_1, \ldots, w_n)$ at the position k iff $k+m-1 \leq n$ and words t_i and w_{k+i-1} match (in the sense introduced above) for $i = 1, \ldots, m$. Term $t_A = (t_1, \ldots, t_n)$ matching some list of words at position a *subsumes* term $t_B = (t_1, \ldots, t_m)$ matching the same list of words at position b iff $a \leq b$ and $b + m \leq a + n$. Term $t_A = (t_1, \ldots, t_n)$ matching some list of words at position a and term $t_B = (t_1, \ldots, t_m)$ matching the same list of words at position b *overlap* iff $a < b < a + n < b + m$.

If term t_A subsumes term t_B, it is almost certainly true that term t_A is more specific. During term extraction we always prefer more specific terms because they are of greater semantic value. For example, Eurovoc term *equality between men and women* subsumes both Eurovoc terms *men* and *women*. Here, the multi-word term is the preferable choice over shorter terms. A less common case is when two terms overlap. For example, phrase *motor vehicle insurance premium* contains two overlapping Eurovoc terms: *motor vehicle insurance* and *insurance premium*.

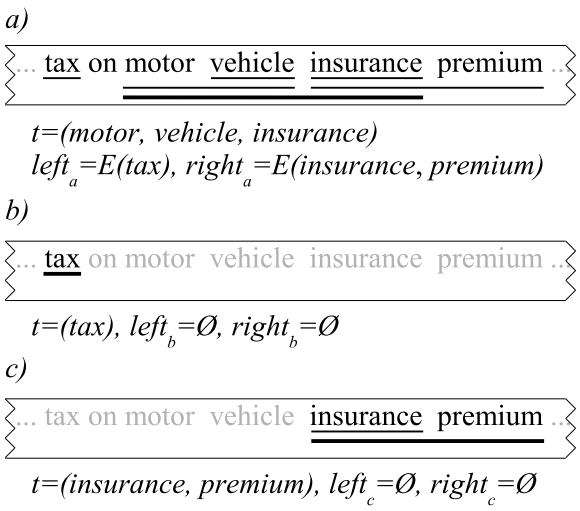

Figure 2: A set of extracted terms contains the longest term extracted in step *a)* and the contents of sets $left_a$ and $right_a$. Sets $left_a = \{(tax)\}$ and $right_a = \{(insurance, premium)\}$ are calculated in steps *b)* and *c)*, respectively.

In these rare cases no term takes precedence over the other, so we choose to extract all of them.

The process of term extraction can be formalised as follows. Let T be a set of all terms, $W^+ = \bigcup_{n=1}^{\infty} W^n$ a set of all word n-tuples and $E : W^+ \rightarrow \wp(T)$ a function mapping a list of words $(w_1, \ldots, w_n) \in W^+$ to a set of extracted terms, element of $\wp(T)$. For example, $E(tax, on, motor, vehicle, insurance, premium) = \{(tax), (insurance, premium), (motor, vehicle, insurance)\}$, as shown in Fig. 2.

Function E is defined recursively as follows. If there are no terms present in the list of words $(w_1, \ldots, w_n)$, then $E(w_1, \ldots, w_n) = \emptyset$. Otherwise, $E(w_1, \ldots, w_n) = \{t\} \cup left \cup right$, where t is the leftmost among the longest terms in the list of words $(w_1, \ldots, w_n)$. If there is no term starting before term t then $left := \emptyset$, otherwise $left := E(w_1, \ldots, w_k)$, where k is the greatest index of ending of all such terms. Set *right* is defined analogously. We choose the leftmost term only to break the tie among the longest terms of the same length. This definition of E ensures that a term that is always subsumed is not extracted, while those that overlap will be extracted.

3.2. Errors in term extraction

The process of term extraction as described above is prone to two kinds of errors: lemmatisation errors and errors due to lexical ambiguity. We continue with a description of these errors, while their relevance is discussed in Section 4.

3.2.1. Lemmatisation errors

Lemmatisation errors may decrease both precision and recall of term extraction. A distinction can be made between lemmatisation failure (given form is not present in the lexicon) and incorrect lemmatisation (given form is related to a wrong lemma).

Since we have $L(w) = \{w\}$ when lemmatisation fails, w_1 and w_2 will match iff $w_1 = w_2$. In other words, if lemmatisation fails on any of the words $(t_1, \ldots, t_m)$ constituting a term, then an exact match for this word is required. This poses no problem if the word in question is not inflective (e.g. functional words, abbreviations etc.). However, if the word is inflective (e.g. nouns, adjectives, verbs, etc.), then various inflective forms of a term, differing from that listed in the thesaurus, will not be found in text. This ultimately leads to a decrease in recall.

If lemmatisation of a word w_1 is incorrect, then it is possible that $L(w_1) \cap L(w_2) \neq \emptyset$ although w_1 and w_2 are in fact not inflections of the same lemma. A single-word term (w_1) will then be incorrectly matched with the word w_2 occurring in text. A mismatch causes a decrease in precision, but it can also cause a decrease in recall if w_1 happens to be an inflective form of another term. It should however be noted that for multi-word terms the probability of this type of error is negligible.

3.2.2. Errors due to lexical ambiguity

Another problem is the lexical ambiguity of natural language, in particular the cases of *homography* and *polysemy*. Homography is a relation between words that have the same orthographic form with unrelated meaning (see Section 2.1). Polysemy refers to the phenomenon of multiple related meanings within a single lemma. Word sense disambiguation is usually performed by examining the context of an ambiguous word. Term extraction as presented in this paper makes no use of disambiguation. Consequently two kinds of errors are possible, both causing a decrease in precision.

First is due to homography: if word w_1 and a single-word term (w_2) are homographs $(L(w_1) \cap L(w_2) \neq \emptyset$ and $L(w_1) \neq L(w_2))$, they will be matched regardless of actual senses in which words w_1 and w_2 are used. Obviously, if w_1 and w_2 are used in different senses, then this is an error. As an example, consider the sentence *"Church bells toll across the town."* Here, the verb *toll* is used in a sense of sounding a bell by pulling a rope, unlike the orthographically identical Eurovoc term, which is a noun, and used in a sense of charge for the use of transport infrastructure. The probability of this kind of error decreases with the number of words constituting a term.

Second kind of error is due to polysemy: if a term is in itself polysemious, matching it to any word in text is always questionable. An example is the sentence *"Sleeping tablet consumption was higher among subjects reporting a bad atmosphere at work."* Here the word *atmosphere* is used in a sense of a surrounding feeling or mood encountered in the working environment, while in the Eurovoc thesaurus it is meant to be used in the sense of physical environment. Again, the more words constitute a term, the less the probability of an error due to polysemy.

4. Experimental results

Eurovoc thesaurus term extraction experiments were conducted on a parallel Croatian-English corpus consisting of 39 legal documents – European Commission Directives and Croatian legislative documents. Eurovoc is a multilingual thesaurus, used by the European Communities, containing over 6000 terms, each of them translated into 21 European languages (including Croatian [2]), thus enabling multilingual information retrieval.

We chose this particular thesaurus because the term extraction described will be a component of a larger indexing system that is using Eurovoc [9].

The number of words ranged from 365 to 26651 for documents in English and from 297 to 19946 for the same set of documents in Croatian. The reason for the larger number of words in English documents lies in the nature of languages, the way documents are translated, but it is mainly due to the existence of articles in English. The number of terms found depends upon the number of words in the text – more terms are found in a longer text. However, we have considered the number of different terms, this being the relevant efficiency parameter of an extraction procedure, since only differing terms carry new information useful for document indexing.

With the method described in the paper, terms were extracted from each Croatian and English document. Then, the same procedure was repeated on the same documents, this time using lemmatisation. Fig. 3 shows the Box and Whisker plot for a total number of extracted terms for English and Croatian, both with and without using lemmatisation.

Wilcoxon matched pairs test for dependent samples confirmed that the difference in the number of extracted terms when using and not using lemmatisation on Croatian documents, as well as when using and not using lemmatisation on English documents, is significant. This implies that the lemmatisation process is very important in the process of term extraction.

Fig. 3 also shows that the effect of lemmatisation on Croatian texts is stronger than on English texts. This was anticipated due to the morphologically rich Croatian language. The effect of lemmatisation is presented in terms of a relative increase in the recall of extracted terms. In our case, the recall is the number of successfully extracted Eurovoc terms divided by the overall number of Eurovoc terms in the text. Since we could not determine the exact number of terms in the text, we computed the difference in the recall relative to the number of extracted terms when the lemmatisation was used. The average increase in recall of the terms for 39 English documents if lemmatisation is used is 0.20, and for Croatian it is 0.53.

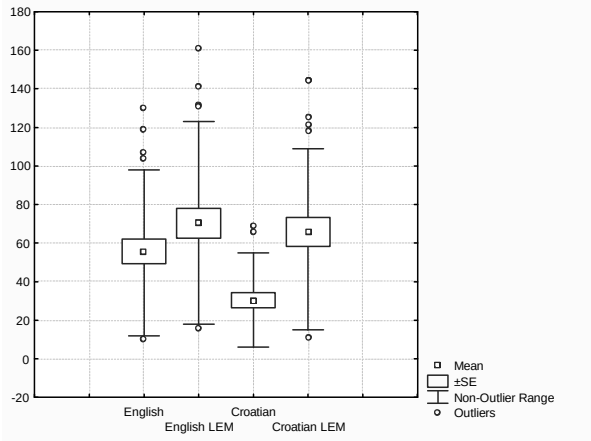

Figure 3: Box and Whiskers plot – number of extracted Eurovoc terms for English and Croatian parallel text, both with and without using lemmatisation.

In Section 3.2 we pointed out that certain kinds of errors can decrease both the recall and precision of term extraction. Decrease of recall is mainly due to lemmatisation failure, which in turn depends on the coverage of the morphological lexicons. We define lexicon coverage as the number of different words in the corpus that were found in the lexicon divided by the total number of different words in the corpus. Lexicons used in our experiments have proven to be of relatively good coverage: 89.71% for English and 92.50% for Croatian lexicon. As for the decrease in precision caused by incorrect lemmatisation, which reflects the linguistic validity of the lexicon, our experiments indicate that these kind of errors were negligible. A hand validation revealed only 0.28% of English and 1.66% of Croatian thesaurus terms to be incorrectly lemmatised. Further decreases in precision due to lexical ambiguity are also negligible: only 1.65% of English as well as Croatian thesaurus terms were mistakenly extracted because of homography and polysemy. Contributing to these low error rates is also the choice of thesaurus terms: e.g. more than 75% of the Eurovoc terms are multi-word terms, and ambiguity of such terms is rare.

5. Conclusion

An enhanced process for thesaurus term extraction has been described and formally defined in the paper. The enhancement was achieved by neutralising the effect of language morphology applying lemmatisation on both the text and the thesaurus terms, and by implementing an efficient recursive algorithm for text extraction.

The statistical evaluation of experimental results has shown that using lemmatisation, term extraction from documents in Croatian is significantly improved, and brought on par to term extraction from documents in English. This clearly indicates that lemmatisation plays an important role in term extraction for highly inflected languages. Furthermore, experiments indicate that errors due to lexical ambiguity in Eurovoc term extraction are rare, making the need for disambiguation methods for this particular thesaurus questionable in practice.

Acknowledgements

We thank another team from Faculty of Electrical Engineering and Computing, Croatian Information Documentation Referral Agency, and Department of Linguistics, Faculty of Philosophy for outstanding work on our joint project Text mining system (2003-082) funded by the Ministry of Science, Education and Sports of the Republic of Croatia. We also thank two anonymous reviewers for helpful comments on an earlier draft of this paper.

References

[1] Automatically Generated Inflection Database. http://wordlist.sourceforge.net [01/16/2005]

[2] Bratanić M, ed. Pojmovnik EUROVOC, 2nd ed., HIDRA, Zagreb; 2000.

[3] Clement L, Sagot B, Lang B. Morphology Based Automatic Acquisition of Large-Coverage Lexica, Proc. of LREC, Lisboa; 2004.

[4] Bakel van B. Modern Classical Document Indexing. Proc. of the 21st ACM SIGIR Conf. on Research and Development in Information Retrieval, Melbourne; 1998.

[5] Croatian National Corpus. http://www.hnk.ffzg.hr [01/16/2005]

[6] Erjavec T, Džeroski S. Machine Learning of Morphosyntactic Structure: Lemmatizing Unknown Slovene Words. Applied Artificial Intelligence, 2004; 18(1): 17–41.

[7] Eurovoc Thesaurus. http://europa.eu.int/celex/eurovoc/ [01/16/2005]

[8] Jackson P, Moulinier I. Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization. Amsterdam: John Benjamins Publishing Co.; 2002.

[9] Kolar M, Vukmirović I, Dalbelo Bašić B, Šnajder J. Computer Aided Document Indexing System. To be published in Proc. of ITI 2005, June 20–23, Cavtat.

[10] NASA Scientific and Tech. Information Program. NASA Thesaurus Machine Aided Indexing. http://mai.larc.nasa.gov/ [01/16/2005]

[11] Nenadić G, Ananiadou S, McNaught J. Enhancing Automatic Term Recognition through Recognition of Variation. Proc. of COLING, Geneva; 2004. p. 604–610.

[12] Oliver A, Tadić M. Enlarging the Croatian Morphological Lexicon by Automatic Lexical Acquisition from Raw Corpora. Proc. of LREC, Lisboa; 2004. p. 1259–1262.

[13] Pouliquen B, Steinberger R, Ignat C. Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. Proc. of the Workshop Ontologies and Information Extraction, Bucharest; 2003.

[14] Ripplinger B, Schmidt P. Automatic Multilingual Indexing and Natural Language Processing. Proc. of SIGIR, Athens; 2000.

[15] Šnajder J. Rule-based automatic acquisition of large-coverage morphological lexicons for information retrieval. Tech. Report, MZOŠ 2003-082, ZEMRIS, FER, University of Zagreb; 2005.