

## Automatic Term Recognition in Polish Texts

The paper describes our experiments in automatic term recognition (ATR) in Polish texts. In particular, we adopt the hybrid C/NC-value method of Frantzi et al. (2000) to identify nested multiword terms in a medical corpus.

Although ATR has been in the research focus for over a decade now, most approaches have been developed for highly positional languages, whereas only a few efforts have been made for Slavic languages which have a richer morphological inflection and a more relaxed word order, e.g., Vintar (2004) (for Slovene) and Nenadic et al. (2003) (for Serbian). In this paper, we report on our experiments in adopting the same method which was used for the other two Slavic languages and applying it to term extraction from Polish texts.

The corpus we used in the experiment consists of 92 diabetes patients' medical records and has been compiled to serve as a lexical resource for developing an information extraction system. The original medical records have been written in Word and, for further processing, they have been converted to a UTF-8 encoded text format. The data contain a few spelling errors but their overall quality is quite good so no further pre-processing was involved. The corpus is untagged and comprises 46449 raw word tokens.

As mentioned above, we adopted the C/NC-value method of Frantzi et al. (2000) which identifies multiword terms as well as subterms they contain. More specifically, C-value measures stability of a candidate term in the corpus by verifying frequency of a term with respect to longer terms in which it is nested. The second measure, NC-value, additionally verifies contexts in which terms appear and ranks higher those which appear in most 'term-friendly' environments (according to Sager (1978), in contrast to non-terms, terms cannot be freely modified; we consider three types of modifiers/contexts: nouns, adjectives and verbs). Both measures rely on a combination of linguistic and statistical techniques: the former is used for the candidate term selection, whereas the latter for ranking.

For the selection of candidate terms, we use linguistic processing in order to identify noun phrases in the texts. In particular, we use a shallow text processing platform, SProUT, which has been integrated with a Polish morphological analyser, cf. Piskorski et al. (2004). We have implemented a few syntactic rules which capture core syntactic structures of Polish NPs: internal NP agreement (a noun has to agree in case, number and gender with a modifying adjective), and a genitive nominal complementation pattern (a noun can take a genitive NP complement). Since SProUT is coupled with a morphological analyser, rather than a tagger, and does not contain a guesser, words which are not in the lexicon remain unrecognized. In fact, many of the unrecognized words are (parts of) terms (e.g., morphological compounds, Latin names of medicines or diseases). After filtering typical 'stop words' (functional words, unattached adjectives, adverbs, verbs, numbers, symbols, punctuation marks, etc.), unknown words are

incorporated to the extracted noun phrases in the post-processing phase. Thus, we group (sequences of) unknown words and either combine them with an adjacent NP or, if there is no adjacent NP, they are treated as a new candidate term. Once a list of candidate terms is created, for each term, its frequency is calculated and contexts relevant for computing NC-value are stored. The final term ranking is performed using the C/NC-value measures.

In the experiment, 2881 candidate terms have been initially extracted. After C/NC-value ranking, this number was reduced to 2762 terms (only terms with non-negative values were considered). We manually evaluated the results and obtained 52% precision and 60.4% recall (recall was measured with respect to frequency of extracted terms rather than all terms which appear in the corpus). The provided results are comparable to those for a similar system for Slovene described in Vintar (2004): 51% precision and 65% recall.

Although the overall performance of the Polish system is quite reasonable, it is not optimal and there is still room for improvement. One of the processing issues is related to imperfect linguistic processing. In particular, since there is no morphological information for unknown words, if they are added to an NP, they can destroy the internal structure of previously identified NPs. For example, in the fragment ‘w [tylnym biegunie] [mikroaneuryzmaty]...’ (in the back pole microaneurismats...), the last word is not initially recognized and it is incorrectly merged with the preceding NP. Also, combining unknown words does not always result in a new term, e.g., ‘glikemia mg’ (glycemy mg) wrongly becomes a term in the post-processing phase (neither of the adjacent words is in the lexicon). In order to remedy this problem, the processing platform should be equipped with a guesser or a tagger. The latter would also help eliminating errors resulting from multiple morphological analyses; for example, the preposition ‘bez’ (without) can be analysed as a genitive plural of the noun ‘beza’ (macaroon) and is often unnecessarily incorporated into candidate terms.

Another factor which affects the performance is related to Polish rich morphology: multiword terms can appear in various inflected forms and each unit in a term can be inflected independently. As a result, frequency counts are skewed. In the present experiment, this issue was partly overcome by using base rather than inflected forms to build candidate terms. Obviously, a better solution must be sought as the true ‘base’ form of a term is usually missed this way (e.g., terms with a genitive NP complement), whereas for unknown words no base form can be currently produced.

#### References:

Frantzi K., Ananiadou S., Mima H. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal of Digital Libraries*.

Nenadic G., Spasic I, and Ananiadou S. (2003). Morpho-Syntactic Clues for Terminological Processing in Serbian. *Proceedings of EACL Workshop on Morphological Processing of Slavic Languages, Budapest, Hungary*.

Piskorski, J., Homola, P., Marciniak, M., Mykowiecka, A., Przepiórkowski, A., Woliński, M. (2004). Information Extraction for Polish Using the SProUT Platform. The proceedings of Intelligent Information Systems 2004 (New Trends in Intelligent Information Processing and Web Mining), Zakopane.

Sager J. C. (1978). Commentary by Prof. Juan Carlos Sager. In Actes Table Ronde sur les Problemes du Decoupage du Termes, Montreal, Quebec, 1978.

Vintar S. (2004) Comparative Evaluation of C-Value in the Treatment of Nested Terms. In Proceedings of the Methodologies and Evaluation of Multiword Units in Real-word Applications, LREC 2004.