Multilingual Approach to e-Learning from a Monolingual Perspective

Vladislav Kuboň and **Miroslav Spousta**
Faculty of Mathematics and Physics
Charles University in Prague
{vk,spousta}@ufal.mff.cuni.cz

Abstract

This paper describes the efforts undertaken in an international research project LT4eL from the perspective of one of the participating languages, Czech. The project aims at exploiting language technologies for adding new functionalities to an open source Learning Management System ILIAS. The new functionalities are based both on existing and newly developed tools for all languages involved. The paper describes in detail the issues of the implementation of a keyword extractor and a glossary candidate detector for Czech.

Introduction

The technological advances in recent years, especially the steadily growing capacity and availability of Internet have stressed the importance of distance learning. The demand for distance education is growing steadily and thus it is no wonder that also the importance of all kinds of tools making the distance education an effective venture for both the teacher and the student is growing.

Language Technology for e-Learning

This situation provided a motivation for an international project called Language Technology for e-Learning (LT4eL). The project aims at several areas where the language technology might be useful, namely at the area of keyword extraction, glossary candidate detection and a semantic search. The new functionalities are being incorporated into an existing open-source system ILIAS (www.ilias.de).

Czech Keyword Extractor

The problem of keyword extraction had been solved for Czech in numerous applications in the past. Probably the oldest implemented solution dates back to 1983 (Kirschner 1983), to a system called MOSAIC (Morphemics Oriented System of Automatic Indexing and Condensation). The system was based on the semantic properties of certain Czech suffixes.

Although this approach provided good results in the past, it turned out to be unsuitable for our project, mainly due to the language and domain dependency of the method.

A Keyword Extraction in LT4eL

A detailed description of the tool for keyword extraction and detection can be found in (Lemnitzer & Degórski 2006). In this paper we would like to present a concise description of its main features.

The important clue whether a given term is a keyword or not, had been introduced by Katz (Katz 1996). According to him, term burstiness is based on the assumption that good keywords tend to appear close to each other in the text. A special attention is being payed to multiword expressions. In most cases the multiword keyword expression consists in Czech of a keyword modified by adjectives, pronouns or numerals and all word forms in the expression agree in gender, number and case. The second important property of multiword keywords is the co-occurrence of the keyword and its dependent modifiers (attributes) in the text, which is usually higher than possible accidental co-occurrence.

Quantitative evaluation of the Czech Keyword Extractor

Our keyword extractor makes use of several different measures of keywordness. First, all possible keyword candidates (including multi-word candidates) are extracted from given document. Then, we apply filtering based on morphosyntactic annotation, filtering uncommon keyword types (e.g. isolated prepositions, conjunctions and numerals). In the second stage we apply scoring function to select the most probable keyword candidates. We use five different scoring functions.

As a base-line, we use TF*IDF ($tfidf$) measure, which is quite common in information retrieval and data mining tasks. In addition $adridf$ (Adjusted Reduced Inverse Document Frequency) measure is used and evaluated (Lemnitzer & Degórski 2006). As many documents used in our project are originally HTML pages, we also extract information from selected HTML elements (headings, bold and italics text, etc.) and mark every possible keyword with these elements.

Another measure we deploy is ARF (Average Reduced Frequency) as defined in (Savický & Hlaváčová 2003). This measure assigns regular corpus frequency to words distributed evenly in the entire corpus, while significantly reducing the frequency for words only occurring in a few narrow clusters. The method is document-boundary indepen-

dent and thus it is not biased by uneven document size. For ranking keywords, we divide regular corpus term frequency by ARF and multiply the result by a term frequency.

We have also tried to combine all these methods using a linear combination computed by the Expectation Maximization algorithm.

Most methods for keyword extraction are directly dependent on the size of corpus used for extraction of background frequency information. Our corpus consists of 46 documents of variable size (1799 – 139 829 tokens), with total size exceeding 1 million tokens. For the purposes of evaluation we manually annotated 18 randomly selected documents. These documents were split into two parts: training (9 documents) and evaluation (9 documents).

In our tests we used methodology described in (Lemnitzer & Degórski 2006), we set up a threshold for all automatic keyword extraction method to 150% of number of manually selected keywords. Results for F_2 measure average across all test documents are shown in Table 1.

Method	tfidf	adridf	Layout	ARF	Combined
F_2	26.21	25.62	24.31	26.84	**28.28**

Table 1: Keyword extractor results (F_2 measure)

Unfortunately, the results are more affected by the nature of documents they were tested upon, they differ much more across documents than they do across individual types of input data, therefore it is impossible to make at the moment any reliable judgment which setup is best. The factor which definitely plays a role is the length of the input document.

Czech Glossary Candidate Detector

The main task of the glossary candidate detection is to find definitory contexts of keywords retrieved by the keyword extractor. The input text has technically a form of an XML text with morphological tags. This input is being processed by simple regular grammars which are different for each language. The grammars are implemented by means of a tool called *lxtransduce* (cf. (Tobin 2005)). This tool belongs to the LTXML2 set of tools developed at the University of Edinburgh.

The Czech training corpus used for GCD has 231 115 tokens in 7 documents on various topics and 1069 hand annotated definitions. The testing corpus consists of 12 smaller documents with 90 571 tokens and 174 hand annotated definitions. Out of the 174 hand annotated definitions 153 are contained in a single sentence, 6 span 2 sentences, and 3 definitions span 3 sentences.

The Czech grammar uses 21 grammar rules at the highest level, divided into 5 groups. The most successful rules belong to the category *NP je/jsou NP* - [NP is/are NP]. Relatively successful are also the rules exploiting certain specific verbs - *definovat* - [define], *znamenat* - [mean], *vymezovat* - [define], *představovat* - [constitute] etc. Table 2 contains the statistics concerning individual types of Czech grammar rules.

The precision and recall of the above mentioned grammar is relatively low: 18.3% and 40.7%, respectively (F_2 measure is 28.9). But the tests of an inter-annotator agreement had shown that not even the human annotators are sure about the definitions either, so the automatic tool generally performs not much worse than the humans.

Pattern	Occurrences in the corpus
NP is/are NP	52
NP verb NP	45
structural	39
NP (NP)	30
NP /:/= NP	20
other patterns	59

Table 2: Czech definition types

The second reason why the detection of definitory contexts scored so low for Czech concerns the typological properties of the language. It has a very rich flection and a relatively very free word order, both properties which are extremely difficult to capture by a simple grammar presented above.

Conclusions

This paper concentrates on the issues encountered during the implementation of language technology tools for one language contained in the international project Language Technology for e–Learning, Czech. The typological properties of Czech enabled to view the problems of language technology tools from the perspective of languages which pose a challenge to many widely accepted natural language processing tools and methods due to the richness of their inflection and the high degree of word-order freedom.

In the future we would like to test some alternative methods for both the keyword extractor and glossary candidate detector. These experiments might answer the crucial question whether by making the language technology tools more language dependent it would be possible to obtain substantially better results for the tasks performed in the project.

References

Katz, S. M. 1996. Distribution of content words and phrases in text and language modelling. *Nat. Lang. Eng.* 2(1):15–59.

Kirschner, Z. 1983. *MOSAIC – A Method of Automatic Extraction of Significant Terms from Texts*. Prague: MFF UK.

Lemnitzer, L., and Degórski, L. 2006. Language technology for elearning – implementing a keyword extractor. Paper presented at the fourth EDEN Research Workshop *Research into online distance education and eLearning. Making the Difference*, 2006 in Castelldefels, Spain.

Savický, P., and Hlaváčová, J. 2003. Measures of Word Commonness. *Journal of Quantitative Linguistics* 9(3):215—231.

Tobin, R. 2005. Lxtransduce, a replacement for fsgmatch. http://www.ltg.ed.ac.uk/richard/ltxml2/lxtransduce-manual.html.