Rule-based Multi-Word Term Extraction, Lemmatization and Description

Cvetana Krstev, Ranka Stanković, Duško Vitas (University of Belgrade)

WG 1, WG2

In this paper we present a rule-based method for multi-word term (MWT) extraction and lemmatization of extracted multi-word terms. Extracted and lemmatized MWT candidates are post-processed using data-driven and heuristic approach in order to reject falsely offered lemmas ("parasite lemmas") and then ranked by calculating various measures before passing them to human evaluators. For accepted terms dictionary entries are automatically produced that enable generation of all terms' inflected forms. All subtasks of this process are integrated into a tool for development and management of lexical resources LeXimir (Stanković et al., 2016).

Motivation

Various approaches have been proposed for MWT extraction that can be grouped into statistical that are based on different statistical measures, rule-based that incorporate linguistic knowledge in various degrees, and hybrid that use both syntactic patterns and statistical measures, mainly for filtering extracted candidates. For highly inflected languages, such as Serbian and other Slavic languages, MWTs can appear in texts in various inflected forms that can affect statistical measures unless these forms are grouped together. For instance, if several inflected forms of a MWT *magnetno polje* 'magnetic field' appear in a text a statistical approach should take into consideration not the frequency of each of them but their total frequency. Moreover, human evaluators would have to repeatedly evaluate many extracted candidates. One possible approach that solves the problem of statistical calculation is to normalize all inflected forms by using a lemmatizer to replace each MWT component by their dictionary lemmas. However, this approach cannot satisfy human evaluators because obtained normalized forms are in many cases grammatically incorrect, as would be the case for our example – the normalized form would be *magnetni*$_{\text{m}}$ *polje*$_{\text{n}}$ in which the agreement of the adjective and the noun in the gender does not hold (because a dictionary lemma of an adjective is always in the masculine gender). Also, from such a normalized form it would not be possible to produce an e-dictionary entry (as requested by Multiflex, see (Savary, 2009)), and consequently all inflected forms.

Solution for MWT extraction, lemmatization and description for Serbian

Corpus preparation and pre-processing. Domain specific texts are collected and processed with e-dictionaries of simple- and multi-word units using Unitex corpora processing system.[1] Besides that, an extract of the Corpus of contemporary Serbian[2] (~ 22 million words) is processed using the same resources for purpose of comparison and calculation of some statistical measures.

MWT extraction. Syntactic patterns in the form of Finite-state transducers (FST) are used on the collection of texts previously implicitly lemmatized and grammatically tagged (but not disambiguated). This enables checking of various agreement and other conditions and results in high-quality recognition. A user can choose among 19 different syntactic structures, corresponding to MWT terms of two to five components that are most frequently used for Serbian MWUs (statistic is computed on the basis of 16,000 MWU lemmas in the Serbian e-dictionary).

MWT lemmatization. Lemmatization is performed in two stages: first a normalized form is obtained by simple-word lemmatization (lemmas are retrieved from e-dictionaries in the 1$^{\text{st}}$ step),

[1] http://www-igm.univ-mlv.fr/~unitex/

[2] http://www.korpus.matf.bg.ac.rs/

and next this form is corrected, if necessary, in order to obtain a MWU lemma. The last step relies on the knowledge of the syntactic structure of the extracted MWT (obtained in the 2nd step) and the e-dictionaries which provide needed component forms. In this step, due to the homography of lemmas and forms, more than one MWU lemma can be offered of which at most one is correct. For instance, *patrona* 'cartridge' and *patron* 'patron' have some identical forms thus leading to the production of two lemmas – *patrona eksploziva* 'explosive cartridge' (correct) and *patron eksploziva* 'explosive patron' (incorrect) – for the extracted form *patronom eksploziva*$_{\text{instrumental}}$.

Parasite lemma rejection. First, the data-driven approach is used. Namely, if several inflected forms of a MWT are retrieved, and for them more than one lemma is offered, then a lemma is chosen that covers the most of the retrieved forms. For instance, from 7 different forms of *patrone eksploziva*, two do not coincide with the forms of the parasite lemma *patron eksploziva* – thus, if some of these discriminative forms are extracted they correctly reject the parasite lemma. If after this step more than one lemma still remains, some heuristic decisions are taken. For instance, if a lemma *patrona eksploziva* with syntactic structure Noun Noun$_{\text{genitive}}$ is offered together with the lemma *patrona eksploziv* (incorrect) with syntactic structure Noun Noun (where two nouns have to agree in the case and the number), then the second one is rejected as being less probable (this structure is used mush less frequently, on the basis of MWU e-dictionary statistics).

Ranking. For evaluation purposes, the user can choose the ranking on the basis of frequency, two association measures T-score and C-Value, and two termhood measures LLR (Log Likelihood Ratio) and Keyness that measure the strength of a MWT compared to some reference source – in our case a reference sample processed in step 1 (Pazienza et al., 2005), (Kilgarriff et al., 2014).

Dictionary production. For evaluated MWTs, and on the basis of the knowledge of their syntactic structure and components lemmas, dictionary entries are automatically produced. For two examples given in this paper these entries are:

magnetno(magnetan.A7:aens1g) polje(polje.N300:ns1q),NC_AXN
patrona(patrona.N600:fs1q) eksploziva,NC_N2X

The information provided by these dictionary entries enables production of all inflected forms associated with values of grammatical categories (gender, number, case, animateness). NC_AXN and NC_N2X are names of FSTs reflecting MWT's syntactic structure that perform this task.

Evaluation results

The whole cycle presented here was performed on corpora of two different domains: library and information science (576,000 words) and mining (more than 625,000 words) while other domains are still in progress. For the first domain, the average precision for retrieval of MWU forms ranged from 0.61 to 0.68, while for the mining, after improving the procedure for the rejection of parasite lemmas, the averaged precision ranged from 0.789 to 0.804. In the latter case, mean average precision of lemma production was 0.95. In our case the calculation of the recall has no point as the performance of our extraction FSTs directly depends on e-dictionary coverage, which is for Serbian e-dictionaries high. The evaluation showed that 94% of distinct multi-word forms were evaluated as proper multi-word units, and among them 97% were associated with correct lemmas.

References

Savary, Agata. "Multiflex: a multilingual finite-state tool for multi-word units." *Implementation and Application of Automata*. Springer Berlin Heidelberg, 2009. 237-240.

Stanković et al. "Rule-based Automatic Multi-Word Term Extraction and Lemmatization". *10th LREC*, 2016

Pazienza, M. T., M. Pennacchiotti, and F. M. Zanzotto. "Terminology extraction: an analysis of linguistic and statistical approaches." *Knowledge mining*. Springer, 2005. 255-279.

Kilgarriff, A. et al. The Sketch Engine: ten years on. *Lexicography* 1(1), 2014. 7-36.