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

Cvetana Krstev, cvetana@matf.bg.ac.rs; Ranka Stanković, ranka@rgf.bg.ac.rs; Duško Vitas vitas@matf.bg.ac.rs;

Human Language Technology group at the University of Belgrade, Studenski trg 1, 11000 Belgrade, Serbia

Dubrovnik, 26-27 September 2016
PARSEME
WG1 WG2

Motivation

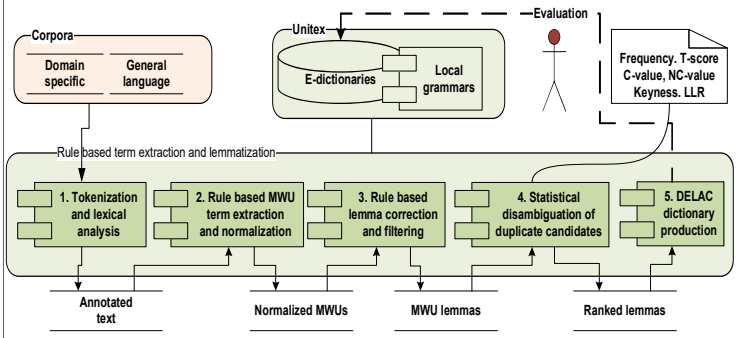
magnetic field

magnetno polje
magnetnog polja
magnetnom polju
magnetnim poljem
magnetna polja
magnetnih polja
magnetnim poljima

$magnetni_m\ polje_n$ → $magnetno_n\ polje_n$

- Gramatically correct lemma important for
  - ✓ human evaluators,
  - ✓ Terminological database,
  - ✓ e-dictionary entry production, and consequently all inflected forms.

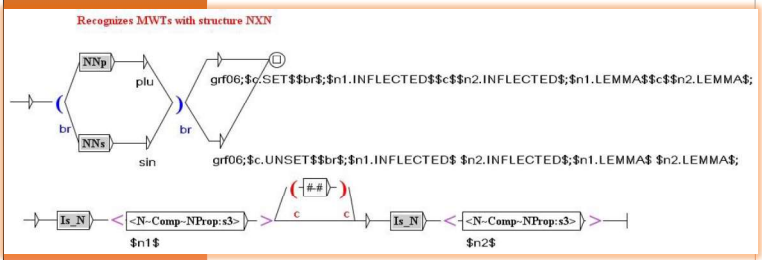
Solution for MWT extraction, lemmatization and description



Corpus preparation and pre-processing

- ✓ Domain specific texts processed with e-dictionaries of simple- and multi-word units using Unitex corpora processing system.
- ✓ Extract of the Corpus of contemporary Serbian (~ 22 M words)

MWT extraction



- ✓ Syntactic patterns in FST form are used on text (lemmatized and grammatically tagged, but not disambiguated).

MWT lemmatization

- ✓ 1st step: normalized form is obtained by simple-word lemmatization (lemmas are retrieved from e-dictionaries),
- ✓ 2nd step: form is corrected, if necessary to obtain a MWU lemma (relies on the syntactic structure of the extracted MWT and the e-dictionaries) - due to homography, more than one MWU lemma can be offered of which only one is correct.

patronom eksploziva

- patrona eksploziva
  'explosive cartridge' (correct ☺)
- patron eksploziva
  'explosive patron' (incorrect ✗)

Automatic lemmatization and description of extracted MWTs

Parasite lemma rejection

- ✓ Data-driven approach: if several inflected forms are retrieved, with more than one lemma, chose one that covers more forms.

patrona eksploziva
patrone eksploziva
patronama eksploziva

patrona eksploziva
~~patron eksploziva~~

- ✓ If more than one lemma still remains, heuristic is applied.

patrona eksploziva → patrona eksploziva (Noun $Noun_{genitive}$)
~~patrona eksplozi~~ (Noun Noun)

Ranking

- ✓ Two unithood 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
- ✓ Combination of unithood and termhood measures

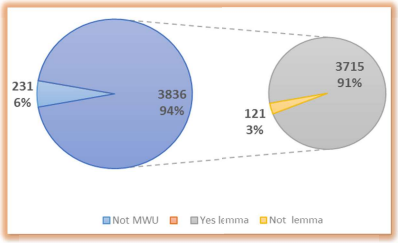
Dictionary production

- ✓ For evaluated MWTs, dictionary entries are automatically produced,
  - ✓ magnetno(magnetan.A7:aens1g) polje(polje.N300:ns1q),NC_AXN
  - ✓ patrona(patrona.N600:fs1q) eksploziva,NC_N2X
- ✓ It enables production of all inflected forms associated with values of grammatical categories (gender, number, case, animateness).

Evaluation results

- ❖ The whole cycle performed on corpora of two different domains, other domains are still in progress.
- ❖ Library and information science (576,000 words), the average precision for retrieval of MWU forms ranged from 0.61 to 0.68
- ❖ Mining (625,000 words), the averaged precision ranged from 0.789 to 0.804. In the latter case, mean average precision of lemma production was 0.95.
- ❖ Recall of our extraction directly depends on e-dictionary coverage, which is for Serbian e-dictionaries high.

❖ **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.

Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić, Aleksandra Trtovac, "Rule-based Automatic Multi-Word Term Extraction and Lemmatization". 10th LREC, 2016