Automatic Detection of Multi-Word Expressions in Lithuanian

STSM report

Justina Mandravickaitė

1 Purpose of the STSM

The aim of this STSM is to explore the possibility of the automatic detection of multi-word expressions (MWE) in Lithuanian via combination of association measures and supervised machine learning. Lithuanian is a synthetic language (favouring morphologically complex words), so simple statistical approaches for MWE cannot provide satisfactory results since this morphological richness results in lexical sparseness. Besides, statistical approaches treat multiword expressions as a bag of words and disregard variation of MWE components. The almost free word order in Lithuanian also does not improve the situation. However, at the moment Lithuanian lexical resources for complementing or replacing statistical approaches are limited. Thus detecting Lithuanian MWE by combining lexical association measures and machine learning is a promising approach in this situation. Machine learning allows various properties of text to be encoded in feature vectors as well as identifying complex non-linear relations. So, it permits capturing elaborate features in languages with complex morphology.

2 Description of the work achieved during the STSM

The research done during the STSM had three phases. The first phase consisted of experiments with small manually annotated corpus of news portal headings (~200 000 words). During the second phase experiments of MWE detection were performed using far larger corpus of transcripted speeches in the Lithuanian Parliament[1] (~ 23 mln. words). Third phase touched exploration of MWE flexibility.

Besides the above activities, I also had the chance to have interesting discussions with members of the Research Group in Computational Linguistics (University of Wolverhampton) and share experience regarding identification and parsing of multiword expressions as it is one of the topics the Group is working on.

2.1 Experiments with the corpus of news portal headings

For experiments two tools there were used: mwetoolkit[2] and WEKA[3]. Effects of MWE identification

1 http://tekstynas.vdu.lt/page.xhtml;jsessionid=5D25C7CB4B5C22ABCCC2385E820D7319?id=projects-current
2 http://mwetoolkit.sourceforge.net/PHITE.php

using different configurations - surface forms only, lemmas, surface forms combined with POS tags, POS patterns – were researched. At the beginning the performance of separate association measures – Loglikelihood, Dice, Maximum Likelihood Estimation, Student's t-score and Poinwise Mutual Information – there was explored. These association measures are implemented in the mwetoolkit. At this phase many inaccuracies in the results provided by morphological tagger (it is rule-based) for Lithuanian there were noticed and thus heavy and time-consuming post-processing in preparation for experiments including morphological information was carried out. Finally, results with values of association measures were used for machine learning experiments in WEKA.

2.2 Experiments with the corpus of transcripted speeches in the Lithuanian Parliament

For experiments the same mwetoolkit and WEKA were used. As POS tagging appeared highly inaccurate and post-processing for correcting mistakes – time consuming, experiments were performed with surface forms only. MWE candidates from 2-grams to 5-grams were extracted. As a reference list MWE candidates with the highest values of association measures there were used. Results with values of association measures combined with evaluation against reference list were used for machine learning experiments in WEKA. Experiments were carried out for MWE in general as well as for special type of MWE – morphological bundles.

2.3 MWE flexibility and variability

For exploring MWE flexibility and variability, Corpus of Contemporary Lithuanian Language (~140 mln. words) was used. Several semi-fixed MWE were extracted in order to explore variability of their flexible part in terms of words and word forms. For calculating variance of the aforementioned semi-fixed MWE Shannon's Diversity Index was used. Calculations were carried out in R[4] (software environment for statistical computing and graphics).

3 Description of main results

During the STSM, all the above activities yielded interesting results, which are described below.

3.1 Automatic MWE identification in the corpus of news portal headings

Experiments with MWE candidates extracted using only surface forms were not successful though mildly interesting. When taking the approach of association measures only, precision, recall & f-measure values were really low. Situation slightly improved with shorter MWE – bi-grams and tri-grams. Association measures in combination with several machine learning algorithms were not able to recognise MWE from manually compiled reference list (annotated manually in the corpus if news headlines) sufficiently. Possibly values of association measures (maximum likelihood estimation, point-wise mutual information,

3 http://www.cs.waikato.ac.nz/ml/weka/
4 https://www.r-project.org/

Student's t-score, dice, log-likelihood) were not so distinguishing in comparison to non-MWE or needed longer corpus & longer "golden list" or/and to add more linguistic information. Situation improved when this approach was tried for bi-grams and tri-grams only. Experiment with lemmatized corpus rendered the values of the association measures mostly equal to zero, with some exceptions of log-likelihood.

Experiments for MWE extraction using POS patterns results were far more satisfying as more correct MWE were identified than using surface forms only. Results were better for shorter MWE (2-4 words). Longer ones seemingly needed more linguistic information as most of them were cultural references and were rare. Also, heavy post-processing was needed in order to correct mistakes of Lithuanian morphological tagger. POS patterns used for successful candidate extraction were the following: Noun + Noun; Verb + Noun; Preposition + Noun.

Approach that combines association measures and supervised machine learning could be defined as successful one. It was applied for MWE candidates extracted using POS pattern information. The “rules” for “True” MWE were extracted using supervised machine learning algorithms. This approach also led to improvement of precision, recall and f-measure values in comparison with just pattern and association measures.

Sadly, reference list of manually annotated MWE was not as useful as expected. It was helpful in exploring certain properties of MWE, e.g., POS patterns, but not so useful for automatic extraction (human and computer logics do not agree?).

3.2 Automatic MWE identification in the corpus of transcripted speeches in the Lithuanian Parliament

Experiments were performed with only part of the corpus as it appeared too big to process with available computer resources. So, out of ~23 mln. words experiments were carried out with ~2 mln. words. Only surface forms were used. Reference list was compiled from candidates with the highest values of association measures and filtered manually. Precision, recall and f-measure for shorter candidates were almost satisfactory, but for longer ones (4- and 5-grams) they were rather low.

Next experiments used bi-gram MWE candidates only. Reference list was taken from the previous experiment, only other MWE than bi-grams were filtered out. Candidates were automatically evaluated against the reference and these results were used for supervised machine learning. Association measures combined with supervised machine learning (eg., JRip algorithm) led to rule extraction for "True" MWE. Precision, recall and f-measure values were higher in comparison with just association measures.

During the last experiment in this phase certain type of MWE was attempted to identify automatically - using association measures in combination with supervised machine learning. Morphological bundles are fixed MWE that are actually grammatical collocations. Even though they are MWE, i.e. word sequence, in morphological analysis they get only 1 POS tag, eg.: *vis dėlto* (nevertheless, though) goes as adverb.

There is a finite list of this type of MWE.

The best results were obtained combining association measures with JRip algorithm. The following "rules" for "True" MWE of this type were automatically extracted:

- (mle_lrs >= 0.000007) and (ll_lrs >= 315.16171) and (ll_lrs >= 984.246322) and (dice_lrs >= 0.101766) and (pmi_lrs <= 7.50661) and (pmi_lrs >= 7.249371)
- (mle_lrs >= 0.000007) and (ll_lrs >= 315.16171) and (ll_lrs >= 982.715054) and (t_lrs <= 16.169638) and (ll_lrs >= 1208.108163) and (pmi_lrs <= 12.192315)
- (mle_lrs >= 0.000007) and (ll_lrs >= 315.16171) and (ll_lrs >= 739.299153) and (ll_lrs <= 767.301909) and (pmi_lrs <= 7.58801)
- (mle_lrs >= 0.000007) and (ll_lrs >= 139.858076) and (ll_lrs >= 552.971325) and (ll_lrs >= 1019.583621) and (ll_lrs <= 1063.205586) and (dice_lrs >= 0.075142)

3.3 Flexibility and variability of MWE

Corpus of Contemporary Lithuanian Language[5] (~140 mln. words) was used for exploring variability and flexibility of several semi-fixed MWE. Example of variability in word usage of flexible part of semi-fixed MWE: *savo kailį* (one's hide (acc.), meaning: life, reputation, good name) *nešti* (to carry)/*gelbėti* (to save)/ saugoti (to protect)/*galvoti apie* (to think about)/*keisti* (change)/*leisti nudirti* (to allow to skin)/*pažinti kaip* (to know as well as)/*sugrįžti į* (to return to)/*parduoti* (to sell)/*išversti* (turn inside out); diversity of this MWE is equal to 1.97841. As for variability in word forms of the flexible part of MWE, there is the following example: *nė su žiburiu + rasti (*not even with the lantern + to find, meaning: not to find something even with the lantern, i.e., to look to something that is not present); flexible part of this MWE (*rasti* – to find*)* in the corpus of Contemporary Lithuanian Language was found in 16 forms, that making diversity of this MWE equal to 2.161354.

3.4 MWE as features for stylometric analysis – additional results

The aim of this experiment was capturing stylistic dissimilarities/variations and mapping positions of the text samples in relation to each other according to gender. Stylometric experiment was performed using corpus of transcripted speeches in the Lithuanian Parliament. Bi-gram MWE were used as features in combination with distance measure and hierarchical clustering. Experiment was a success as feature vector from 50 to 1200 bi-gram MWE was able to show difference in speech according to gender in the Lithuanian Parliament.

5 http://tekstynas.vdu.lt/

4 Future collaboration with host institution

We intend to continue our work carried out during the STSM. Future collaboration will likely include continued work on automatic identification of MWE, flexibility and variability analysis as well as MWE application to stylometric analysis.

5 Foreseen publications resulting from the STSM

We plan to submit papers on the topics of the STSM, for instance, to compare MWE flexibility in Lithuanian and English. Also, we are working on a joint paper together with dr. Michael Oakes on application of MWE as features in computational stylistics.

6 Confirmation by the host institution of the successful execution of the STSM

I confirm that Justina's STSM visit to Wolverhampton was entirely successful, and was of benefit to Justina, myself, and our Research Group in Computational Linguistics as a whole. Justina managed to complete all the work originally set out in the STSM application, as well as an additional study in computational stylometry. As planned, she applied the MWE characterization measures of flexibility and diversity developed by Hanks et al. to a free word-order language (Lithuanian) for the first time. She also used a number of standard measures of collocational strength (both on the Ramisch package and using machine learning techniques) to discover idioms in her own corpus of Lithuanian newspaper headlines and a corpus of Lithuanian Parliamentary Proceedings. The additional study on computer stylometry found that it was possible to detect differences in gender based on differential usage of MWEs among Lithuanian politicians. We hope that the work on stylometry will form the basis of a published paper, and the other work will be presented as a poster at the PARSEME meeting in Skopje. Justina took part in many useful discussions with the members of our research group, several of whom are working on MWEs. She also attended a number of group seminars.