Automatic Extraction of Grammatical Multiword Expressions for Lithuanian (WG3)

Justina Mandravickaitė
Baltic Institute of Advanced Technology, Lithuania
Vilnius university, Lithuania
justina@bpti.lt

Michael P. Oakes
University of Wolverhampton, United Kingdom
Michael.Oakes@wlv.ac.uk

Tomas Krilavičius
Baltic Institute of Advanced Technology, Lithuania
Vytautas Magnus University, Lithuania
t.krilavicius@bpti.lt

1 Introduction

In this abstract we discuss an experiment in the automatic extraction of grammatical multiword expressions (MWEs) for Lithuanian. This could provide more insight into text semantics, and in this way improve different NLP tasks. We believe that the results of this research will contribute to the further explorations of MWEs in Lithuanian.

2 Grammatical Multiword Expressions

Grammatical MWEs became of interest to linguists during the annotation of the Lithuanian corpus as it was not possible to annotate certain words separately (Kovalevskaitė, 2012). These words are quite common in contemporary Lithuanian and indicate a certain phraseological tendency of the language. Grammatical MWEs have been discussed in Kovalevskaitė (2012), Kovalevskaitė and Rimkutė (2009). They are defined as fixed expressions consisting of two or more functional words (inflected or non-inflected) that have a unified common meaning, are non-compositional and also have a syntactic function (Rimkutė, 2009; Rimkutė and Kovalevskaitė, 2010). They stand for adverbs, prepositions, conjunctions, particles, pronouns and interjections, e.g.: *bet kur* – anywhere, *iš naujo* – anew, *iš esmės* – [get] to the point, *be perstojo* – continuously, *norom nenorom* – willynilly (to do something though unwilling), etc.

3 Corpus of Transcribed Lithuanian Parliamentary Speeches

We have chosen the corpus of transcribed Lithuanian parliamentary speeches for our experiments[1]. It contains speeches of members of the Lithuanian Parliament (MPs) from March 1990 to December 2013. The number of MPs is 147, i.e. only MPs with at least 200 speeches. The minimum number of words in an individual speech is 100. The size of the whole corpus is 23,908,302 words (Kapočiūtė-Dzikienė*et al.*, 2014).

Due to the limited computational resources, for the extraction of grammatical MWEs only 2,530,445 words were used in our experiments, i.e., the last parliamentary term speeches (2008-2012).

4 Method

We used lexical association measures (LAMs) combined with supervised machine learning algorithms in this investigation. The first part of the experiment was executed with mwetoolkit[2] (Ramisch, 2015) and the second one - using the WEKA[3] (Hall *et al.*, 2009) implementation of selected machine learning algorithms.

Firstly, using mwetoolkit, the candidate MWE bi-grams were extracted from the raw text. Then values of 5 association measures (Maximum Likelihood Estimation, Dice's coefficient, Pointwise Mutual Information, Student's t score and Log-likelihood score) (Ramisch, 2015) were calculated. Afterwards preliminary results were evaluated against the reference list of bi-gram grammatical MWEs selected as in Rimkutė (2009). The list consisted of 335 grammatical MWEs.

In the second part (using WEKA), preliminary results were evaluated against the reference list of bi-gram grammatical MWE (converted to ARFF file with the values of True (it is MWE) and False (it is not MWE)). Several algorithms (Naïve Bayes, Bayesian Network and Random Forest) were applied for automatic extraction of MWEs. As the data was rather sparse we also separately used two filters – SMOTE (it resamples a dataset by applying the Synthetic Minority Oversampling TEchnique) (Chawla *et al.*, 2002) and Resample (it produces a random subsample of a dataset using

[1] Developed for the project "Automatic Authorship Attribution and Author Profiling for the Lithuanian Language" (acronym ASTRA) (No. LIT-8-69)

[2] http://mwetoolkit.sourceforge.net/PHITE.php

[3] http://www.cs.waikato.ac.nz/ml/weka/

either sampling with replacement or without replacement) (Hall *et al.*, 2009).

5 Results

We conducted the experiments with 317 grammatical MWEs out of the 335 presented in the reference list found in the transcribed parliamentary speeches.

The results of our experiments in different scenarios (lexical association measures only, lexical association measures combined with a supervised machine learning algorithm, lexical association measures combined with a supervised machine learning algorithm and one of the filters – SMOTE or Resample) are presented in Table 1.

	Precision	**Recall**	**F-meas.**
LAMs	0.14%	**95%**	0.29%
LAMs + Bayesian Networks	3.4%	37.2%	6.1%
LAMs + Naïve Bayes with SMOTE	5.1%	25.1%	8.5%
LAMs + Random Forest with Resample	**94.4%**	**62.2%**	**75%**

Table 1.Summary of the results.

Using only the lexical association measures implemented in the mwetoolkit combined with the reference list for evaluation recall[4] was 95%, but precision[5] (0.14%) and F-measure[6] (0.29%) were very low, i.e. it seems that almost any candidate MWEs out of the 219,900 candidates was identified as an MWE. Thus, association measures do not suffice for the successful extraction of grammatical MWEs for Lithuanian.

Combining association measures and supervised machine learning algorithms we used 3 scenarios: without any filter, with the SMOTE filter and with the Resample filter. All the models were tested using standard 10-fold cross-validation. The best results without any filter were achieved with the Bayesian Network classifier (118/317 correct MWEs). Using SMOTE the best results were achieved with the Naive Bayes classifier (159/317 correct MWEs) and using the Resample filter – with the Random Forest classifier (204/317 correct MWEs).

Hence, combining association measures with supervised machine learning improves extraction of grammatical MWEs for Lithuanian.

6 Conclusions and Future Plans

We report our experiment for extraction of grammatical MWEs for Lithuanian by combining lexical association measures and supervised machine learning. This experimental setup improved our results in comparison with using association measures only. Our future plans include experiments for automatic extraction of different types of MWEs for Lithuanian and a greater diversity of MWEs.

Acknowledgments

This research was partially funded by COST (Parseme, IC1207).

References

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 321-357.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, Witten I. H. 2009. The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Kapočiūtė-Dzikienė, J., Utka, A., and Šarkutė L. 2014. Seimo posėdžių stenogramų tekstynas autorystės nustatymo bei autoriaus profilio sudarymo tyrimams. Linguistics: Germanic & Romance Studies/Kalbotyra: Romanu ir Germanu Studijos 66.

Kovalevskaitė, J. 2012. Lietuvių kalbos samplaikos. Ph.D. thesis, Vytauto Didžiojo universitetas.

Kovalevskaitė, J. and Rimkutė, E. 2009. Morfologininių samplaikų struktūros ypatumai: kelių kalbų palyginimas. Darbai ir Dienos 50: 120-156.

Ramisch, C. 2015. Multiword expressions acquisition: A generic and open framework. Theory and Applications of Natural Language Processing series XIV, Springer.

Rimkutė, E. 2009. Gramatinė morfologinių samplaikų klasifikacija. Kalbų studijos 14: 32-38.

Rimkutė, E. and Kovalevskaitė, J. 2010. Sudėtinės ir suaugtinės lietuvių kalbos morfologinės samplaikos. Kalbų studijos 16: 79-88.

[4] Recall is the proportion of grammatical MWEs returned by the algorithm and present in the reference list.

[5] Precision is the proportion of correct grammatical MWEs in all the results returned by algorithm.

[6] F-measure is a harmonized average of Precision and Recall.