Identification of Multiword Expressions in Parallel Latvian and Lithuanian Corpus (WG3)

[1,2]Justina Mandravickaitė, [1,3]Tomas Krilavičius and [4]Inguna Skadina
[1]Baltic Institute of Advanced Technology (LT)
[2]Vilnius University (LT),
[3]Vytautas Magnus University (LT),
[4]University of Latvia (LV)
justina@bpti.lt, t.krilavicius@bpti.lt, inguna@latnet.lv

Abstract

We discuss an experiment on automatic identification of bi-gram multiword expressions in parallel Latvian and Lithuanian corpora. Our approach uses raw corpora and combination of lexical association measures and supervised machine learning. We have achieved 92,4% precision and 52,2% recall for Latvian and 95,1% precision and 77,8% recall - for Lithuanian.

1 Introduction

We report experiments on automatic detection of multi-word expressions (MWEs) in Latvian and Lithuanian. Both languages are synthetic, i.e. favor morphologically complex words, thus detecting MWEs by combining lexical association measures (LAMs) and machine learning (ML) could be a right approach.

ML allows encoding properties in feature vectors (lexical, morphological, syntactic, semantic, contextual, etc.) associated with classes, as well as identify non-linear relations and capturing elaborated features in morphologically rich languages.

LAMs compute an association score for each collocation candidate by assessing the degree of association between its components. Combining LAMs helps in the collocation extraction task (Pecina, 2008; Pecina and Schlesinger, 2006; Pecina, 2010), even by combining a relatively small number of measures. So far there is no universal combination of LAMs that works best, since collocation extraction depends on the data, language and type/notion of MWEs.

2 Method

We combine LAMs and ML. LAMs were obtained with *mwetoolkit*[1] (Ramisch, 2015) and for ML for MWEs candidates with LAMs values WEKA[2] (Hall et al., 2009) was used. The candidate MWE bi-grams were extracted with *mwetoolkit* from the raw text and 5 association measures (*Maximum Likelihood Estimation, Dice's coefficient, Pointwise Mutual Information, Student's t score* and *Log-likelihood score*) were calculated. The reference lists based on EuroVoc - Multilingual Thesaurus of the European Union[3] were used for evaluation. Then *Naïve Bayes* (John and Langley, 1995), *OneR* (rule-based classifier; (Holte, 1993)) and *Random Forest* (Breiman, 2001) were applied.

SMOTE (applies *Synthetic Minority Oversampling TEchnique* to resample dataset) (Chawla et al., 2002) and Resample (it produces a random subsample of a dataset using either sampling with replacement or without replacement) filters were used due to the sparseness (Hall et al., 2009).

Precision, Recall and F-measure (Powers, 2011) were used to evaluate the results.

1/3 Latvian and Lithuanian parts of *JRC-Acquis Multilingual Parallel Corpus* (Steinberger et al., 2006)[4], i.e. $\sim$ 9 mln. words for each language were used. As there are no known *gold standards* MWE evaluation resources for Latvian and Lithuanian, to evaluate MWE candidates with calculated LAMs, we used EuroVoc, a Multilingual Thesaurus of the European Union. We selected bi-gram terms only, as statistical methods were reported to be more successful with shorter n-grams (Bartsch and Evert, 2014). We used separate MWE's lists for Latvian (3608 bi-grams) and Lithuanian (Lithuanian - 3783).

We chose surface forms as our experience showed that lemmata produced zero values for most LAMs. Features are numerical vectors of LAMs values combined with booleans TRUE (MWE) and FALSE (not MWE). The latter values are obtained after evaluation of candidate MWEs against reference list. For the next stage we plan to use GIZA++ translation probability scores of MWE candidates as features as well.

[1] *mwetoolkit*, http://mwetoolkit.sourceforge.net
[2] http://www.cs.waikato.ac.nz/ml/weka/
[3] EuroVoc, the EU's multilingual thesaurus, http://eurovoc.europa.eu/drupal/
[4] *JRC-Acquis Multilingual Parallel Corpus*, https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis

	Scenario	Precision	Recall	F-meas.
LV	LAMs	0.1%	21.4%	0.3%
	LAMs+NaiveBayes	0.6%	4.3%	1.1%
	LAMs+OneR+SMOTE	**100%**	13.3%	23.4%
	LAMs+Random Forest+Resample	92.4%	**52.2%**	**66.7%**
LT	LAMs	0.2%	19.4%	0.2%
	LAMs+NaiveBayes	0.6%	4.6%	1.1%
	LAMs+OneR+SMOTE	**100%**	12.6%	22.4%
	LAMs+RandomForest+Resample	95.1%	**77.8%**	**85.6%**

Table 1: Summary of the results for Latvian (LV) and Lithuanian (LT)

3 Experiments and Discussions

We performed experiments with 736 MWE present in the corpus from the reference list for Lithuanian and with 772 for Latvian. See experimental results for LAMs only, LAMs+ML algorithm, LAMs+ML algorithm+SMOTE/Resample in Table 1.

Using only the lexical association measures implemented in the *mwetoolkit* combined with the reference list for evaluation gave low results. Especially low was Precision - 0.1% for Latvian and 0.2% for Lithuanian. Thus it seems that almost any candidate MWE out of the 558 772 (Latvian) and 587 406 (Lithuanian) was identified as an MWE. Thus, association measures did not suffice for the successful extraction of MWEs for Latvian and Lithuanian.

LAMs and ML algorithms were combined in 3 ways: (i) without any filter, (ii) with the SMOTE filter and (iii) with the Resample filter. 10-fold cross-validation was used.

The best results for Latvian were achieved with Random Forest classifier and the Resample filter ($P = 92.4\%, R = 52.2\%$ and $F = 66.7\%$). The best results for Lithuanian were achieved with same configuration, reaching $P = 95.1\%, R = 77.8\%$ and $F = 85.6\%$.

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

4 Conclusion

We report our experiment for extraction of MWEs, that is, bi-gram terms for Latvian and 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 Latvian and Lithuanian and a greater diversity of MWEs as well as exploration of additional features for better results, e.g. GIZA++ probability scores.

References

S. Bartsch and S. Evert. 2014. Towards a firthian notion of collocation. *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2nd Work Rep. of the Acad. Net. Internet Lexicography.*

L. Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Jrnl. of Artif. Int. Res.*, pages 321–357.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Expl.*, 11(1):10–18.

R. C. Holte. 1993. Very simple classification rules perform well on most commonly used datasets. *Machine learning*, 11(1):63–90.

G. H. John and P. Langley. 1995. Estimating continuous distributions in bayesian classifiers. In *Proc. of the 11th conf. on Uncertainty in Artificial Intelligence*, pages 338–345.

P. Pecina and P. Schlesinger. 2006. Combining association measures for collocation extraction. In *Proc. of the COLING/ACL*, pages 651–658. ACL.

P. Pecina. 2008. *Lexical Association Measures: Collocation Extraction.* Ph.D. thesis, Charles U.

P. Pecina. 2010. Lexical association measures and collocation extraction. *Language resources and evaluation*, 44(1-2):137–158.

D. M. Powers. 2011. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

C. Ramisch. 2015. *Multiword expressions acquisition: A generic and open framework.* Theory and App. of Natural Language Processing.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv, cs/0609058.*