Identification of Multiword Expressions for Latvian and Lithuanian: Hybrid Approach

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

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

Abstract

We discuss an experiment on automatic identification of bi-gram multiword expressions in parallel Latvian and Lithuanian corpora. Raw corpora, lexical association measures (LAMs) and supervised machine learning (ML) are used due to deficit and quality of lexical resources (e.g., POS-tagger, parser) and tools. While combining LAMs with ML is rather effective for other languages, it has shown some nice results for Lithuanian and Latvian as well. Combining LAMs with ML 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 explore applicability of the automatic detection of multi-word expressions (MWEs) in Latvian (LV) and Lithuanian (LT). Both languages belong to Baltic language group and are synthetic (favor morphologically complex words), thus simple statistical approaches for identification of MWEs do not provide satisfactory results, as the morphological richness leads to lexical sparseness. Representations, such as bag of words ignore variation of MWEs components (Sharoff, 2004). The relatively free word order in both languages does not improve the situation. Lexical resources for complementing or replacing statistical approaches are limited. However, exploration of MWEs flexibility and morpho-syntactic rules could improve detection of MWEs in Lithuanian easier. But even most of the hybrid methods cannot be implemented in a straightforward manner due to limited availability of lexical resources and tools, e.g. POS tagger, parser, etc.

Thus possibility of detecting Latvian and Lithuanian MWEs by combining lexical association measures and machine learning could be a right approach in this situation. Machine learning allows various properties of text to be encoded in feature vectors (lexical, morphological, syntactic, semantic, contextual, etc.) associated with output classes, as well as identifying complex non-linear relations. It permits capturing elaborate features in languages with complex morphology.

2 Combining LAMs and Supervised Machine Learning

Combination of *lexical association measures* (LAMs) and supervised machine learning algorithms is already under scrutiny, (Zilio et al., 2011) use it for the extraction and evaluation of MWEs from the English part of *Europarl Parallel Corpus*, extracted from the proceedings of the European Parliament; (Dubremetz and Nivre, 2014) explores extraction of nominal MWEs from the French part of the Europarl corpus using application of the same method. Performance of different combinations of LAMs is discussed in (Pecina and Schlesinger, 2006; Pecina, 2008a; Pecina, 2008b; Pecina, 2010).

LAMs compute an association score for each collocation candidate assessing the degree of connection between its components. Scores can be used for the extraction of collocation candidates, ranking and classification (rejecting collocations below (above) threshold).

Different groups of collocations differ in sensitivity to certain association measures depending on their types, e.g., collocations where components statistically occur more often than incidentally, *Log-likelihood ratio*, x^2 *test*, *Odds ratio*, *Jaccard*, *Pointwise mutual information* perform better, while for collocations occurring in the different contexts than their components (non-

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 97–101,
Valencia, Spain, April 4. ©2017 Association for Computational Linguistics

compositionality principle) *J-S divergence*, *K-L divergence*, *Skew divergence*, *Cosine similarity* in vector space are preferred suggested (Pecina, 2008b). For discontinuous MWE (with other words in amidst the components of MWE), *Left context entropy* and *Right context entropy* perform better (Pecina, 2008b).

Combining association measures, even a relatively small number, helps in the collocation extraction task (Pecina, 2008a), (Pecina and Schlesinger, 2006), (Pecina, 2010), however there is no the best universal combination of association measures, since the task of collocation extraction depends on the corpora, language and type/notion of MWEs.

3 Experimental Setup

We use LAMs combined with supervised machine learning. LAMs are calculated using *mwetoolkit*[1] (Ramisch, 2015), and WEKA[2] (Hall et al., 2009) is used to train selected classifiers LAMs.

In this paper we disccuss experiments with bi-gram MWEs only, but we plan to extended definitions of LAMs to tri- and tetra-grams, which is not always straighforward, and explore LAMs+ML approach for longer MWE in future research.

Candidate MWE bi-grams were extracted from the raw text with *mwetoolkit*: frequencies of separate words and bi-grams are counted, hapaxes are removed, and values of 5 association measures (*Maximum Likelihood Estimation, Dice's coefficient, Pointwise Mutual Information, Student's t score* and *Log-likelihood score*) (Ramisch, 2015) are calculated. For each language, the results were evaluated against the reference lists, based on EuroVoc - Multilingual Thesaurus of the European Union[3].

The results were evaluated against the reference list of bi-gram MWE (converted to ARFF file with the values of **true** (MWE) and **false** (not MWE)) using WEKA. Selected algorithms (*Naïve Bayes* (John and Langley, 1995), *OneR* (rule-based classifier; (Holte, 1993)), *Bayesian Network* (Su et al., 2008) and *Random Forest* (Breiman, 2001)) were applied for automatic identification of MWEs. Feature vectors were constructed from LAMs values for each MWE candidate and its appearance in reference list (**true/false**).

SMOTE (it re-samples a dataset by applying the *Synthetic Minority Oversampling TEchnique*) (Chawla et al., 2002) and Resample (it produces a random subsample of a dataset using either sampling with or without replacement) (Hall et al., 2009) filters were used to deal with data sparseness.

To **evaluate** performance we employ (i) *precision* $P = \frac{tp}{tp+fp}$, (ii) *recall* $R = \frac{tp}{tp+fn}$ and (iii) *F-score* $F_1 = 2 \cdot \frac{P \cdot R}{P+R}$, where tp, fp and fn are *true positives* (correctly identified MWEs), *false positives* (expressions incorrectly identified as MWEs) and false negatives (incorrectly identified as non-MWEs), correspondingly (Powers, 2011; Perry et al., 1955).

Association measures and supervised machine learning algorithms were combined in 3 ways: (i) without any filter, (ii) with the SMOTE filter and (iii) with the Resample filter. All the models were tested using standard 10-fold cross-validation.

4 Corpus and Reference Source

4.1 Corpus

1/3 of Latvian and Lithuanian parts of *JRC-Acquis Multilingual Parallel Corpus* (Steinberger et al., 2006)[4], containing the total body of European Union law applicable to its member states (selected texts written since 1950s), i.e., $\sim$ 9 mil. words for each language, were used. Preprocessing consisted of tokenizing (one sentence per line) and lowercasing only, because the goal is to get the best possible results without relying on special linguistic tools, e.g., POS tagger, parser.

4.2 Reference Source for Evaluation of MWE Candidates

As there was known *gold standard* MWE evaluation resources for Latvian and Lithuanian, we use bi-grams from EuroVoc (a Multilingual Thesaurus of the European Union). We use separate lists for each language to evaluate MWE candidates with calculated LAMs values, resulting in `.arff` file with numerical values of LAMs and logical values showing, whether record is **true** (MWE) and **false** (not MWE). Latvian reference list consists of 3608 bi-gram terms, while Lithuanian list has 3783 bi-gram items. Number of bigrams was different, because MWEs in Lithuanian/Latvian not

[1] `http://mwetoolkit.sourceforge.net`

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

[3] EuroVoc, the EU's multi-lingual thesaurus, `http://eurovoc.europa.eu/drupal/`

[4] `https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis`

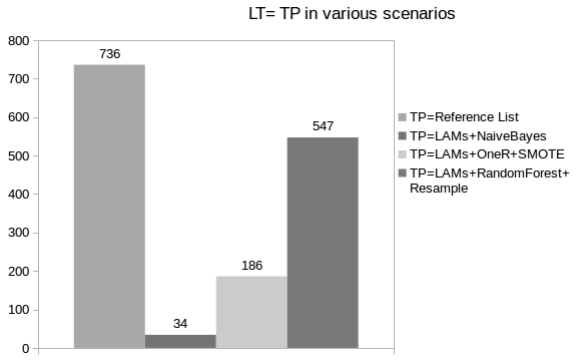

Figure 1: Lithuanian TP in various scenarios

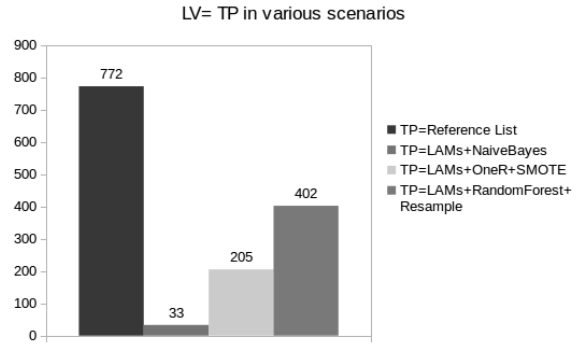

Figure 2: Latvian TP in various scenarios

always had their equivalents as bi-grams in other language and vice versa, e.g. coal - *akmens anglys* (Lithuanian), *akmeņogles* (Latvian); pasture fattening - *ganomasis gyvulių penėjimas* (Lithuanian), *nobarošana ganībās* (Latvian)

5 Results

We experimented with 736 (LT) and 772 (LV) MWEs present in the corresponding corpus from the reference. See Figures 1 and 2 for results, Table 1 for summary of experimental results (LAMs only, LAMs combined with a supervised machine learning, LAMs combined with a supervised machine learning and filters).

Referece list was based on EuroVoc which mostly contained the EU institutions related terms, hence MWEs mostly fitted into 3 categories: Noun + Noun, Adjective + Noun and Abbreviation or Acronym + Noun. However, as we did not use either POS tagger or parser (see the beginning of the paper), detailed morpho-syntactic analysis is in our future plans.

Using only the lexical association measures implemented in the *mwetoolkit* against the reference, performance was low: $R = 21.4\%$ and 19.4%, and $P = 0.1\%$ and 0.2%, and $F_1 = 0.3\%$ and 0.2%, for LV and LT, respectively. Almost any candidate MWE out of the 558 772 (LV) and 587 406 (LT) was identified as an MWE. Thus, association measures did not suffice for the successful extraction of MWEs for Latvian and Lithuanian.

The best results for Latvian without any filter were achieved with the Naïve Bayes classifier (33/772 correct MWEs), reaching P=0.6%, R=4.3% and F1=1.1%.

Using SMOTE the best results were achieved with the OneR classifier (205/772 correct MWEs; $P = 100\%, R = 13.3\%$ and $F_1 = 23.4\%$) and using the Resample filter – with the Random Forest classifier (402/772 correct MWEs with $P = 92.4\%, R = 52.2\%$ and $F_1 = 66.7\%$).

The best results for Lithuanian without any filter were achieved with the Naïve Bayes classifier (34/736 correct MWEs with $P = 0.6\%, R = 4.6\%$ and $F_1 = 1.1\%$). Using SMOTE the best results were achieved with the OneR classifier (186/736 correct MWEs, having $P = 100\%, R = 12.6\%$ and $F_1 = 22.4\%$) and using the Resample filter – with the Random Forest classifier (547/736 correct MWEs; we reached $P = 95.1\%, R = 77.8\%$ and $F_1 = 85.6\%$).

Results show, that combining LAMs with supervised ML improves extraction of MWEs for both languages.

6 Analysis of Misclassified MWE Candidates

Configuration LAMs + Random Forest + Resample performed best for both languages. However, there were misclassified MWE candidates and below there is a more detailed analysis of errors made by Random Forest classifier.

6.1 False Positives

For Lithuanian 22 unique items were misclassified as MWEs and for Latvian - 31 (sampling was done with replacement, thus some items were repeated). False positives belong to one of 3 groups of errors (see Table 2):

(i) good candidates for MWE, but not present in the EuroVoc, and thus not included in the reference list (e.g., LT: *augimo stimuliatorius* (growth stimulator), *traktorių konstrukcijos* (tractor constructions); LV: *valsts slieksnis* (national threshold), *valsts tiesībās* (state law)); (ii) error, occurred due to low frequency (2-3); (iii) real False Positive

Scenario	Precision	Recall	F-meas.
Latvian			
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%**
Lithuanian			
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 and Lithuanian

Latvian	
MWE, not in EuroVoc	6
Low frequency	18
Debatable MWE candidates	7
Lithuanian	
MWE, not in EuroVoc	6
Low frequency	8
Real false positives	7

Table 2: Summary of False Positives for Latvian and Lithuanian

or debatable MWE candidate that needs confirmation.

6.2 False Negatives

For Lithuanian 132 unique items were misclassified as non-MWEs and for Latvian - 336 (sampling was done with replacement, thus some items were repeated). False negatives belong to one of 2 groups of errors (see Table 3):

(i) error, occurred due to extremely low frequency (2-3); (ii) error, occured due to relatively low frequency (3-10). For most misclassified items in the group of extremely low frequency there were pairs of MWE candidates with the same LAMs values (e.g., LT: *vertikalusis susitarimas* & *valdybų susitarimas* (vertical agreement & board agreement); LV: *vispārējais budžets* & *vispārējais labums* (general budget & overall benefit)). Low frequency group mostly had unique combinations of LAMs values.

Results show that heavier filtering according to frequencies should be considered, e.g., filtering out candidates with < 20 occurrences (Evert, 2008). Beside frequency, other LAMs have to be taken into consideration as there is a possibility

Latvian	
Very low frequency (2-3)	109
Low frequency (3-10)	227
Lithuanian	
Very low frequency (2-3)	47
Low frequency (3-10)	85

Table 3: Summary of False Negatives for Latvian and Lithuanian

that *Maximum Likelihood Estimation, Dice's coefficient, Pointwise Mutual Information, Student's t score* and *Log-likelihood score* were not capable to capture all the properties of MWE candidates correctly.

7 Conclusions

We report our experiment for extraction bi-gram MWEs for Latvian and Lithuanian by combining lexical association measures and supervised machine learning. This method appears to be more effective for Lithuanian than Latvian. All in all, using ML together with LAMs improved results: the best configuration LAMs + Random Forest + Resample filter achieved $F_1 = 66.7\%$ for Latvian and $F_1 = 85.6\%$ for Lithuanian. However, an exception was the second-best configuration LAMs + OneR + SMOTE, where results for Latvian were slightly better ($F_1 = 23.4\%$) than for Lithuanian ($F_1 = 22.4\%$).

Future plans include further analysis of low frequency MWEs, because it was a reason for a significant number of errors. Exploration of other LAMs could help to deal with it, and correctly capture complexities of Latvian and Lithuanian. Using EuroVoc is a poor man's solution, us-

ing it resulted in getting a high number of False Positives, which seem to be good candidates for MWEs. Of course, it would be interesting to move from bi-grams, to tri- and tetra-grams as well.

Acknowledgments

This research was funded by a grant (No. LIP-027/2016) from the Research Council of Lithuania.

References

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

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, pages 321–357.

Marie Dubremetz and Joakim Nivre. 2014. Extraction of nominal multiword expressions in french. *EACL 2014*, page 72.

Stefan Evert. 2008. A lexicographic evaluation of german adjective-noun collocations. *Towards a Shared Task for Multiword Expressions (MWE 2008)*, page 3.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

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

George H John and Pat Langley. 1995. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 651–658. Association for Computational Linguistics.

Pavel Pecina. 2008a. *Lexical Association Measures: Collocation ExtractionLexical Association Measures: Collocation Extraction.* Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic.

Pavel Pecina. 2008b. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 54–61. Citeseer.

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

James W Perry, Allen Kent, and Madeline M Berry. 1955. Machine literature searching x. machine language; factors underlying its design and development. *American Documentation*, 6(4):242–254.

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

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

Serge Sharoff. 2004. What is at stake: a case study of russian expressions starting with a preposition. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 17–23. Association for Computational Linguistics.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.

Jiang Su, Harry Zhang, Charles X Ling, and Stan Matwin. 2008. Discriminative parameter learning for bayesian networks. In *Proceedings of the 25th international conference on Machine learning*, pages 1016–1023. ACM.

Leonardo Zilio, Luiz Svoboda, Luiz Henrique Longhi Rossi, and Rafael Martins Feitosa. 2011. Automatic extraction and evaluation of mwe. In *8th Brazilian Symposium in Information and Human Language Technology*, pages 214–218.