Automatic Identification of Multi-Word Expressions for Latvian and Lithuanian

Justina Mandravickaitė, Tomas Krilavičius, Ka Lok Man

***Abstract*— We discuss an experiment on automatic identification of bi-gram multiword expressions for Latvian and Lithuanian. As these languages are considered to be underresourced in terms of lexical resources and availability or accuracy of special lexical tools (e.g. POS-tagger, parser), 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..**

***Index Terms*—hybrid approach, lexical association measures, machine learning, multi-word expressions.**

I. Introduction

A Multi-Word Expression (MWE) is a sequence of >=2 words, which functions as a single unit at linguistic analysis, e.g. syntactic analysis. Identification of MWEs is one of the most challenging problems in NLP. Many techniques are used for this problem, however, not all of them can be transferred to Lithuanian and Latvian.

Latvian and Lithuanian languages belong to the Baltic language group and are synthetic languages (favor morphologically complex words), thus simple statistical approaches for identification of MWEs cannot provide satisfactory results, as the morphological richness results in lexical sparseness.

Statistical approaches which treat multiword expressions as a bag of words pay no attention to the variation of MWE components [15]. The relatively free word order in both languages also does not improve the situation. Moreover, Lithuanian and Latvian lexical resources for complementing or replacing statistical approaches are limited.

However, exploration of MWEs flexibility and adding exceptions could make the detection of MWE in Lithuanian easier. But even most of the hybrid methods cannot be implemented in a straightforward manner. 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.

Combination of lexical association measures (LAMs) and supervised machine learning algorithms was investigated by several authors, e.g. [17] used such approach for the extraction and evaluation of MWEs from the English part of Europarl Parallel Corpus, extracted from the proceedings of the European Parliament; extraction of nominal MWEs by application of the same method and from the French part of the same Europarl corpus is reported by [18]. Best combinations of LAMs are extensively reported in [9], [8], [11], [10].

LAMs compute an association score for each collocation candidate assessing the degree of association between its components. These scores can be used for the extraction of collocation candidates, ranking them, or for classification (setting a threshold and dismissing all collocations below it). However, some association measures are very similar (e.g., Pointwise Mutual Information and Dice identify lexical collocations; T-score and Loglikelihood show grammatical collocations [4]).

Different subgroups of collocations have different sensitivity to certain association measures depending on their extraction principle. For example, for 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 which occur in the different contexts than their components (non-compositionality principle) J-S divergence, K-L divergence, Skew divergence, Cosine similarity in vector space were suggested [10]. For discontinuous MWE (where other words occur among the components of MWE), Left context entropy and Right context entropy show better results [10].

Combining association measures helps in the collocation extraction task [9], [8], [11]. Improvement of the extraction procedure can be achieved by combining a relatively small number of measures. And so far there is no universal combination of association measures that works best in every situation, since the task of collocation extraction depends on the data, language and type/notion of MWEs.

II. Methodology

We used lexical association measures (LAMs) combined with supervised machine learning algorithms in this investigation. The first part of the experiment (getting values of LAMs) was executed with mwetoolkit[1] [14] and for the second one (application of machine learning algorithms for MWEs candidates with LAMs values)

J. Mandravickaitė is with Baltic Institute of Advanced Technology and Vilnius University, Lithuania (corresponding author to provide phone: +370 683 87737, e-mail: justina@bpti.lt).

T. Krilavičius is with Baltic Institute of Advanced Technology and Vytautas Magnus University, Lithuania (e-mail: t.krilavicius@bpti.lt).

K. L. Man is with Xi'an Jiaotong-Liverpool University, China (e-mail: kalok2006@gmail.com).

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

ISBN: 978-988-14047-7-0
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

WEKA[2] [5] was used.

Firstly, using mwetoolkit, the candidate MWE bi-grams were extracted from the raw text. Then, values of 5 association measures (Maximum Likelihood Estimation (mle), Dice's coefficient (dice), Pointwise Mutual Information (pmi), Student's t score (t) and Log-likelihood score (ll)) [14] were calculated. Afterwards preliminary results were evaluated against the reference lists of bi-gram MWE for each language. The aforementioned reference lists were based on EuroVoc - Multilingual Thesaurus of the European Union[3].

Aftewards, preliminary 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 [7], OneR (rule-based classifier; [6]), and Random Forest [2]) were applied for automatic identification of MWEs.

As the data was rather sparse we separately used two filters: SMOTE (it re-samples a dataset by applying the Synthetic Minority Oversampling TEchnique) [3] and Resample (it produces a random subsample of a dataset using either sampling with replacement or without replacement) [5].

The evaluation of classification results were based on standard measures - Precision, Recall and F-measure. As in [13], [12], Precision is the proportion of items, predicted by supervised machine learning algorithm, which are relevant to the query; Recall is proportion of items, predicted by supervised machine learning algorithm, which are relevant to the query and are predicted successfully. F-measure can be defined as the average of Precision and Recall when they are close, and in general it is the square of the geometric mean divided by the arithmetic mean in terms of the aforementioned Precision and Recall [13].

We have chosen Latvian and Lithuanian parts of JRC-Acquis Multilingual Parallel Corpus[4] [16]. It contains the total body of European Union law applicable to its member states. Currently it includes selected texts written since 1950s. Statistics for Latvian (LV) and Lithuanian (LT) parts of JRC-Acquis Multilingual Parallel Corpus are presented in Table 1.

TABLE 1
LATVIAN (LV) AND LITHUANIAN (LT) PART OF JRC-ACQUIS MULTILINGUAL PARALLEL CORPUS

Characters	Words	Language
196 452 051	27 594 514	LT
199 438 258	26 967 773	LV

We used 1/3 of each, Latvian and Lithuanian, parts of JRC-Acquis Multilingual Parallel Corpus, i.e. 9 mln. words for LV and LT each.

Our purpose was to get the best possible results without relying on special linguistic tools, e.g. POS tagger, parser. Thus preprocessing consisted of tokenizing (one sentence per line) and lowecasing only.

As there are no known gold standard MWE evaluation resources for Latvian and Lithuanian at the moment, to evaluate MWE candidates with calculated LAMs, extracted with mwetoolkit, we used EuroVoc, a Multilingual Thesaurus of the European Union. We selected bi-gram terms only, as statistical methods were generally reported to be more successful with shorter n-grams [1]. We used separate lists (one for Latvian, one for Lithuanian) of these bi-gram MWEs for evaluation of MWE candidates with calculated LAMs values. We have got an .arff file which, beside numerical values of LAMs, included logical values, showing, whether record is True (MWE) and False (not MWE). Latvian reference list consisted of 3608 bi-gram terms, while Lithuanian reference list had 3783 bi-gram items.

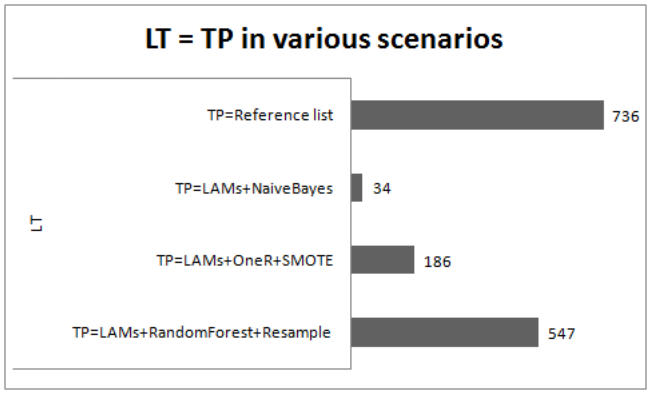

Fig. 1. Lithuanian TP in various scenarios.

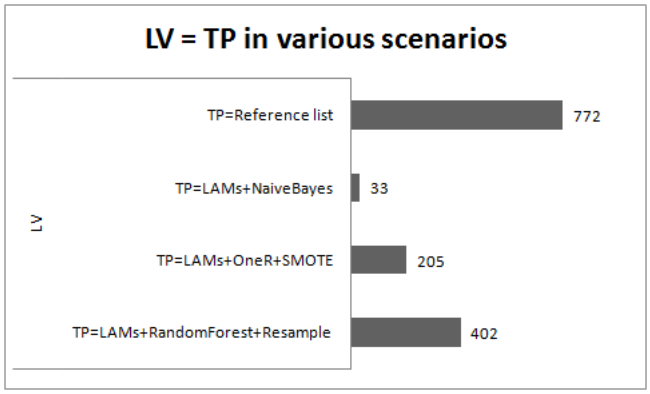

Fig. 2. Latvian TP in various scenarios.

III. Experiment

We performed experiments with 736 MWE present in the corpus from the reference list for Lithuanian, that is, 736 true positives (TP). For Latvian there were 772 compounds present in the corpus from reference list, i.e. we had 772 MWEs. For TP in different scenarios, see Figure 1 and Figure 2. Summary of experimental results performed in different scenarios (LAMs only, LAMs combined with a supervised machine learning algorithm, LAMs combined with a supervised machine learning algorithm and one of the filters – SMOTE or Resample) are presented in Table 2.

[2] http://www.cs.waikato.ac.nz/ml/weka/
[3] http://eurovoc.europa.eu/drupal/
[4] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis

ISBN: 978-988-14047-7-0
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE 2
SUMMARY OF THE RESULTS FOR LATVIAN AND LITHUANIAN

F-measure	Recall	Precision	Scenario
Latvian			
0.3%	3.5%	0.2%	LAMs
1.1%	4.3%	0.6%	LAMs + Naïve Bayes
23.4%	13.3%	**100%**	LAMs + OneR + SMOTE
66.7%	**52.2%**	92.4%	LAMs + Random Forest + Resample
Lithuanian			
2.2%	4.9%	1.4%	LAMs
1.1%	4.6%	0.6%	LAMs + Naïve Bayes
22.4%	12.6%	**100%**	LAMs + OneR + SMOTE
85.6%	**77.8%**	95.1%	LAMs + Random Forest + Resample

Using only the lexical association measures implemented in the mwetoolkit combined with the reference list for evaluation gave low results. Recall was 3.5% for Latvian and 4.9% - for Lithuanian. Precision was 0.2% for Latvian and 1.4% for Lithuanian. Finally, F-measure was 0.3% and 2.2% for Latvian and Lithuanian respectively. These results were observed after several gradual frequency filtering, setting collocation boundaries via LAMs value curves (see Table 3) and adjustments in terms of range of candidate MWEs. Out of all 5 LAMs, relative frequency or mle measure proved to be nearly useless in our case. Thus in LAMs scenario 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 in our case.

TABLE 3
THRESHOLDS OF LAMS VALUES FOR LATVIAN AND LITHUANIAN

Lithuanian	Latvian	LAMs
0,123	0,018	dice
336,774	126,549	ll
0,000	0,000	mle
7,742	7,632	pmi
11,179	7,452	t

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.

The best results for Latvian without any filter were achieved with the Naïve Bayes classifier (33/772 correct MWEs), reaching 0.6% for Precision, 4.3 % - for Recall and 1.1% - for F-measure. It was probably too simple for capturing and predicting complicate relations between MWE candidate components. Still, without any filter, these were the best results. Using SMOTE the best results were achieved with the OneR classifier (205/772 correct MWEs; 100% for Precision, 13.3% - for Recall and 23.4% - for F-measure) and using the Resample filter – with the Random Forest classifier (402/772 correct MWEs with 92.4% of Precision, 52.2% of Recall and 66.7% of F-measure).

The best results for Lithuanian without any filter were achieved with the Naïve Bayes classifier (34/736 correct MWEs with 0.6% of Precision, 4.6% of Recall and 1.1% of F-measure). The results showed that this classifier performed even worse then LAMs only. Again, it was probably too simple for capturing and predicting complicate relations between MWE candidate components. Still, without any filter, these were the best results. Using SMOTE the best results were achieved with the OneR classifier (186/736 correct MWEs, having 100% for Precision, 12.6% - for Recall and 22.4% - for F-measure) and using the Resample filter – with the Random Forest classifier (547/736 correct MWEs; we reached 95.1% of Precision, 77.8% of Recall and 85.6% of F-measure).

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

IV. Conclusions

We report our experiment for extraction of MWEs, that is, bi-gram terms for Latvian and Lithuanian. Because of the lack of lexical resources and availability or accuracy of special lexical tools (e.g. POS-tagger, parser), we used raw corpora and combination of 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 Lithuanian and a greater diversity of MWEs.

References

[1] 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.

[2] L. Breiman. 2001. Random forests. Machine learning, Vol. 45, No. 1, pp. 5–32,.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. Jrnl. of Artif. Int. Res., Vol. 16, pp. 321–357.

[4] S. Evert and B. Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. Computer Speech & Language, 19(4):450–466.

[5] 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., Vol. 11, No. 1, pp. 10–18.

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

[7] 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.

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

[9] P. Pecina. 2008a. Lexical Association Measures: Collocation Extraction, Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic.

[10] P. 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.

ISBN: 978-988-14047-7-0
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

[11] P. Pecina. 2010. Lexical association measures and collocation extraction. Language resources and evaluation, 44(1-2):137–158.
[12] J. W Perry, A. Kent, and M. M. Berry. 1955. Machine literature searching x. machine language; factors underlying its design and development. American Documentation, 6(4):242–254.
[13] D. M. Powers. 2011. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.
[14] C. Ramisch. 2015. Multiword expressions acquisition: A generic and open framework. Theory and App. of Natural Language Processing.
[15] S. 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.
[16] 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.
[17] L. Zilio, L. Svoboda, L. H. L. Rossi, and R. M. Feitosa. 2011. Automatic extraction and evaluation of mwe. In 8th Brazilian Symposium in Information and Human Language Technology, pages 214–218.
[18] M. Dubremetz and J. Nivre. 2014. Extraction of nominal multiword expressions in French. EACL 2014, page 72.

ISBN: 978-988-14047-7-0
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)