MULTIPLE FEATURES FOR MULTIWORD EXTRACTION: A LEARNING-TO-RANK APPROACH

Elena Tutubalina (tutubalinaev@gmail.com), Kazan Federal University, Kazan, Russia
Pavel Braslavski (pbras@yandex.ru), Ural Federal University, Yekaterinburg, Russia

This paper describes the extraction of multiword expressions (MWEs) from corpora for inclusion in a large online lexical resource for Russian. The novelty of the proposed approach is twofold: 1) we use two corpora – the Russian National Corpus and Russian Wikipedia – in parallel and 2) employ an extended set of features based on both data sources. To combine syntactic and statistical features derived from two corpora, we experiment with several learning-to-rank (LETOR) methods that have been proven to be highly effective in information retrieval (IR) scenarios. We make use of bigrams from existing dictionaries for learning, which leads to very sparing manual annotation efforts. Evaluation shows that machine-learned rankings with rich features significantly outperform traditional corpus-based association measures and their combinations. Analysis of resulting lists supports the claim that multiple features and diverse data sources improve the quality of extracted MWEs. The proposed method is language-independent.

Key words: multiword expressions (MWEs), collocations, lexical acquisition, learning-to-rank methods (LETOR), thesaurus, Russian language

ИЗВЛЕЧЕНИЕ МНОГОСЛОВНЫХ ВЫРАЖЕНИЙ НА ОСНОВЕ МНОЖЕСТВЕННЫХ ПРИЗНАКОВ И МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ РАНЖИРОВАНИЮ

Тутубалина Е.В. (tutubalinaev@gmail.com), Казанский федеральный университет, Казань, Россия
Браславский П.И. (pbras@yandex.ru), Уральский федеральный университет, Екатеринбург, Россия

Ключевые слова: многословные выражения, устойчивые словосочетания, машинное обучение ранжированию, тезаурус, русский язык

1. Introduction

Multiword expressions (MWEs) are heavily underrepresented in existing Russian lexical resources. We encountered the problem of MWE extraction within a project aimed at creating a new wordnet for Russian. The study described in the paper deals with nominal bigrams – the most common MWE type. Since the pioneering work by Church and Hanks (1989) the problem of MWE extraction has been studied in depth, and various statistical association measures (AMs) have been proposed. Despite the task seems to be solved, larger datasets and advanced statistical methods available nowadays offer opportunities for a more efficient solution.

The proposed approach includes three components: 1) we use two different corpora – the Russian National Corpus (364M tokens) and Russian Wikipedia (1.2M articles, 318M tokens) – in parallel; 2) MWE candidates are described with a rich set of features (various corpus-based statistics, link-based Wikipedia features, phrase structure, Web statistics, etc.); 3) we formulate the MWE ranking task in terms of multiple 'queries' and 'documents' and apply learning-to-rank (LETOR) algorithms that showed good results in information retrieval (IR) scenario. Our approach deals with different kinds of MWEs – collocations, idioms, set phrases, etc. (see classification in Baldwin and Kim (2010)) – in a uniform way. We took several thousands of nominal bigrams from existing Russian dictionaries and manually labeled them as positive and negative examples. This routine allowed us to minimize manual labeling efforts and is more advantageous than such alternatives as labeling output of an automatic method, which can potentially introduce bias towards presented results, or asking an expert to produce a list of good and bad examples from scratch, which is very labor-intensive. Using limited training data, we were able to rank the whole set of candidate MWEs extracted from both corpora and cut them off at desired level (our estimate of the target number of MWEs for the wordnet under development is around 40K). Evaluation showed that proposed approach outperformed existing AMs, as well as classification-based methods. Manual probes proved that high-ranked MWEs are good enough to be included in the resource with minimal manual intervention. The method is language-independent – it relies only on the availability of a large corpus, Wikipedia, and a part-of-speech (POS) tagger. Furthermore, the method is highly flexible and can be applied to other MWE types.

2. Related Work

There is a large body of literature on extraction of multiwords, collocations, and keyphrases; Hasan (2014) and Ramisch (2015) provide an extensive overview of the field. Three groups of approaches related to our work can be distinguished: (i) methods based on purely statistical AMs; (ii) machine-learned classification; (iii) Wikipedia-based approaches to terminology extraction.

Traditional approaches rank a list of MWEs according to their co-occurrence frequencies or statistical AMs (Evert and Krenn, 2005; Pecina and Schlesinger, 2006). Krenn and Evert (2001) evaluated Mutual Information (MI), Dice coefficient, Student's t-score and log-likelihood ratio for adjective-noun pairs. Pecina and Schlesinger (2006) evaluated 82 measures on a Czech corpus. Some studies suggested different strategies for handling low-frequency and high-frequency items (Evert and Krenn, 2001; Evert and Krenn, 2005; Bouma, 2009). Wermter and Hahn (2006) showed that the most advanced AMs perform similarly to raw frequency.

State-of-the-art studies consider the MWE extraction task as a classification problem (Pecina and Schlesinger, 2006; Fothergill and Baldwin, 2011; Karan et al., 2012; Ramisch, 2015). Pecina and Schlesinger (2006), Ramisch et al. (2010) and Karan et al. (2012) employed support vector machines (SVM) with frequency counts, traditional AMs, and POS patterns as features. These supervised approaches are different from ours in that Karan et al. (2012) and Ramisch (2015) created a training set consisting of positive and negative MWE examples, while Fazly and Stevenson (2007) and Fothergill and Baldwin (2011) assigned MWE categories. Feature-rich ranking of keyphrases extracted from a document is close to our approach (Jiang et al., 2009). However, extracting keyphrases from a document exploits quite a different set of document-level features such as position of the first occurrence, document field (e.g. title, section heading, anchor text), and text highlighting (e.g. boldface). Document-level keyphrase extraction task differs from our setting in that the same word sequence occurring in different documents can be a good keyphrase in one case, but not suitable in other cases.

Many studies explored Wikipedia as an external knowledge resource for terminology extraction (Hartmann et al., 2012; Vivaldi et. al., 2012) and keyphrase extraction (Medelyan et al., 2009). Medelyan et al. (2009) used a machine learning approach with Wikipedia-based semantic features to determine whether the document can be annotated with a given keyphrase. Hartmann (2012) considered n-grams that appeared in Wikipedia titles and anchor text as candidates for subsequent ranking by AMs. (Vivaldi et al., 2012) used Wikipedia categories to validate term candidates extracted from scientific texts.

Due to limited space we do not survey a large body of literature on learning to rank and feature selection for IR; Liu (2009) gives a nice overview of approaches and methods. In our work we follow the feature selection approach proposed by Geng et al. (2007) that combines two scores: importance of individual features and similarity between features.

3. Data

In our study, we use two corpora – Russian National Corpus[1] (RNC) and Russian Wikipedia[2]. RNC has genre subdivisions – scientific texts, classical literature, legal and official documents, religious texts, children's literature, nonfiction, news, etc. – that we use for feature calculation. We treat Wikipedia both as a "plain text corpus" to calculate MWE statistics and as

[1] http://ruscorpora.ru/en/

[2] http://ru.wikipedia.org

semi-structured data: we make use of Wikipedia links, redirects, categories, and page titles. Lemmatization and POS-tagging is performed with *mystem* library[3].

We consider all bigrams conforming to one of six morpho-syntactic patterns – *Adjective + Noun, Noun + Adjective, Participle + Noun, Noun + Participle, Noun + Noun (genitive),* and *Noun + Noun (instrumental)* – as candidate MWEs. Moreover, a candidate MWE must occur at least ten times in the RNC or to be a Wikipedia title.

We also collected nominal bigram entries from three dictionaries: Wiktionary[4] (3,155), Small Academic Dictionary (2,955), and Ushakov's Dictionary (2,506), which resulted in 7,751 unique bigrams in total. Manual inspection revealed that the list contained many archaisms (e.g. *книга живота – book of life*, *духовное брашно – spiritual repast*), narrowly used metaphorical expressions (e.g., *деревянный макинтош – coffin* (literally – *wooden mackintosh*), *белый друг – toilet bowl* (*white friend*)), joking expressions (e.g., *губозакаточная машинка – lip-rolling machine*), as well as named entities (e.g. *Амурская область – Amur Region*). The list underwent manual labeling by two lexicographers. Lexicographers labeled MWEs as positive (suitable for a general-purpose thesaurus) and negative (otherwise). Manual labeling resulted in an approximately equal number of positive and negative examples. Table 1 summarizes the data used in the study.

# of positive examples	3,981
# of negative examples	3,770
# of unique words in labeled examples	5,871
# of positive examples in the test set	1,322
# of candidate MWEs from the RNC	190,416
# of candidate MWEs from Wikipedia	157,748
# of unique candidate MWEs from both corpora	329,866
# of candidate MWEs overlapping with labeled set (RNC)	82,456
# of candidate MWEs overlapping with labeled set (Wiki)	117,837
# of unique candidate MWEs overlapping with labeled set	188,441

Table 1. Candidate MWEs and labeled data (overlapping bigrams have at least one common word).

Most advanced LETOR algorithms (so-called pair-wise and list-wise methods, see (Liu, 2009)) optimize ranking in the context of individual queries and respective result lists in contrast to earlier point-wise approaches that model relevance as global regression or classification task.

[3] https://tech.yandex.ru/mystem/

[4] http://ru.wiktionary.org

In order to apply modern LETOR algorithms to the MWE extraction task, we represent the data as a set of “queries” and “documents”. Our hypothesis is that ‘divide and conquer’ approach helps deal with MWEs of different types and frequency ranges in a unified way in the learning phase. For “queries”, we took 5,871 unique words from labeled examples to create individual lists of MWE candidates (“documents”) containing the “query” (see Table 2). 56.5% of all candidates were included at least in one list. We randomly sampled 80% of the ‘queries’ for training and held out 20% for testing.

word (‘query’)	**overlapping bigrams (‘documents’)**
неправильный	неправильная установка (wrong installation), неправильная постановка (wrong statement), неправильная музыка (wrong music), неправильная галактика (wrong galaxy), неправильная переменная (wrong variable), неправильная дробь (improper fraction)
струна	слабая струна (weak string), натянутая струна (tense string), гетеротическая струна (heterotic string), бозонная струна (bosonic string), квантовая струна (quantum string), золотая струна (gold string), космическая струна (cosmic string), спинная струна (notochord)
корова	белая корова (white cow), старая корова (old cow), черная корова (black cow), синяя корова (blue cow), священная корова (sacred cow), дойная корова (milk cow), морская корова (sea cow)
вещество	специальное вещество (special substance), обычное вещество (usual substance), рабочее вещество (working substance), солнечное вещество (solar substance), сухое вещество (solid), белое вещество (white substance), мягкое вещество (soft substance), полярное вещество (polar substance), компактное вещество (compact substance), радиоактивное вещество (radioactive material), живое вещество (live substance), лекарственное вещество (medicinal substance), действующее вещество (active ingredient), вредное вещество (harmful substance), серое вещество (gray substance), простое вещество (simple substance), органическое вещество (organic), химическое вещество (chemical agent)

Table 2. ‘Queries’ (single words from labeled bigrams) and ‘documents’ (overlapping candidate MWEs); positive examples are underlined.

4. Methods

To apply a ranking algorithm to the data we have to present each MWE candidate as a feature vector. Note that, in the IR scenario a vector represents a query-document pair, i.e. there are features depending on the query, document, or both. In our case, all features describe an individual MWE independently from the “query”, which allows us to apply the obtained ranking function later to the global set of candidates (hundreds of thousands items). The feature set used in the study (42 features in total) is described below.

RNC features (14): RNC global frequency, ten frequencies in genre subcorpora (reflects specificity of the MWE), first and second words’ frequencies, the presence of the candidate in the corpus.

Wikipedia-based features (20) included: Wikipedia frequency, the presence of a redirect with the given MWE, match with a Wikipedia title, the number of in- and out-links, the number of categories assigned to the page, the presence of an infobox, 11 binary features corresponding to the infobox type[5], and capitalization (the latter three features aimed at capturing named entities).

Structural features (7) included six binary features corresponding to the above mentioned extraction patterns plus bigram length in characters (indirectly reflects the bigram specificity).

Web document frequency (1) refers to the number of documents returned to MWE as a phrase query by a search engine (SE) through an API[6].

We used three algorithms implemented in the RankLib library[7] to obtain MWE rankings: MART (Friedman, 2001), RankBoost (Freund et al., 2003), and LambdaMART (Wu et al., 2007) with default parameters. To improve efficiency of the training, we applied a feature selection (Geng et al., 2007). We held out 20% of the training set as validation set to optimize the number of features. First, according to the method, we computed importance of each feature using *mean reciprocal rank* (MRR). We measured similarity between features with Kendall's τ for pairs of corresponding rankings. Second, we maximized the sum of the importance scores of individual features and minimized the total similarity score between the features using a greedy search algorithm. Finally, five groups of features with the best results on the validation set were used to evaluate LETOR models on the test set.

5. Evaluation

We evaluated multiple intermediate rankings with artificial queries using two measures: 1) *mean reciprocal rank* (MRR) and 2) *bpref*, an evaluation measure suited for incomplete judgments (Buckley and Voorhees, 2004). MRR is an average of inverse ranks of the first positive example in each ‘query’; while *bpref* accounts for inversions – cases, when ‘relevant’ items are ranked lower than ‘non-relevant’ ones. Both measures were averaged over 1,449 lists in the test set.

We compared our approach to state-of-the-art collocation extraction methods based mainly on frequency (Pecina and Schlesinger, 2006; Ramisch et al., 2010; Karan et al., 2012). In particular, we implemented the best-performing method for keyphrase extraction (Jiang et al., 2009) based on SVM-rank[8] algorithm and following features: POS patterns, MWE frequency, and 20 AMs calculated using UCS toolkit[9] on (i) RNC, (ii) Wikipedia, (iii) both corpora. We

[5] http://en.wikipedia.org/wiki/Wikipedia:List_of_infoboxes

[6] https://xml.yandex.ru/

[7] http://sourceforge.net/p/lemur/wiki/RankLib/

[8] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

[9] http://www.collocations.de/software.html

also implemented AMs (t-score, log-likelihood, and MI) as baselines. Evaluation results are presented in Table 3.

Ranking method	**MRR**	***bpref***
MI	0.440	0.353
t-score	0.615	0.321
log-likelihood	0.620	0.353
Wikipedia frequency	0.625	0.467
RNC frequency	0.624	0.328
SVM-rank (RNC)	0.644	0.550
SVM-rank (Wikipedia)	0.609	0.492
SVM-rank (RNC+Wikipedia)	0.635	0.483
MART	0.639	0.545
MART + feature selection	0.639	0.480
LambdaMART	0.679	0.742
LambdaMART + feature selection	0.684	0.546
RankBoost	0.739	0.742
RankBoost + feature selection	**0.758**	**0.825**

Table 3. MRR and *bpref* measures computed on the test set

As the results show, LambdaMART and RankBoost scored best compared to MART, SVM-Rank and AMs. SVM-rank and MART scores are comparable. Impact of feature selection is mixed: it improved both MRR and *bpref* for RankBoost, but degraded LambdaMART and MART *bpref* scores. Best LambdaMART results were obtained with all features except for the Wikipedia title feature and four Wikipedia infobox features. RankBoost scored best using the Wikipedia title feature, number of categories, and presence of candidate in the corpus. Table 4 illustrates the contribution of different feature groups to the overall performance. The results support our initial hypothesis that multiple data sources improve results.

	MRR highest	**MRR lowest**
all features	0.679	0.598
w/o RNC-based features	0.565	0.497
w/o Wikipedia-based features	0.609	0.543
w/o structural features	0.671	0.592
w/o results from the search engine	0.678	0.602

Table 4: MRR for highest and lowest positive items ranked with LambdaMART: contribution of different feature groups.

Top-40K lists ranked by LambdaMART, SVM-Rank, and RNC-based frequency contain 634, 472, and 452 positive examples (out of 990 ‘relevant’ MWEs in the initial global list), respectively. In the top-40K MWEs ranked by LambdaMART, 43% items occur in both corpora, 35% and 22% occur in Wikipedia or RNC only, respectively. This again illustrates the benefit of using two data sources in parallel. Figure 1 presents ROC curves for the top-40K candidate MWEs ranked by LambdaMART, SVM-Rank, and RNC-based frequency (note that the total number of true positives differs for these 40K-lists, see above). Table 5 shows MWEs at different levels of the global list ranked by LambdaMART.

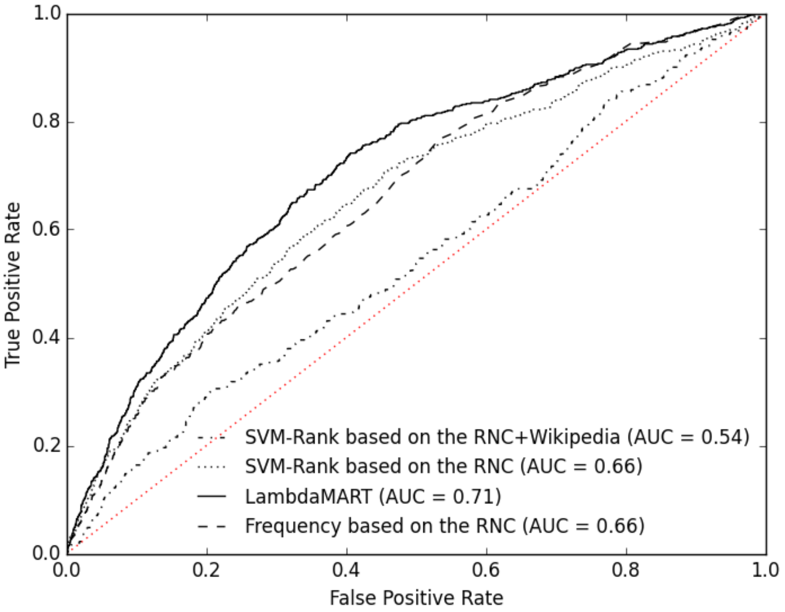

Fig. 1: ROC curves for four methods.

Cut-off level = 100	**Cut-off level = 1,000**
земная кора (Earth’s crust)	подсадная утка (decoy-duck)
программное обеспечение (software)	народный дух (national character)
основные фонды (basic assets)	разговорная речь (spoken language)
биологические науки (bioscience)	публичная библиотека (public library)
общественное мнение (public opinion)	братская могила (mass grave)
Cut-off level = 2,500	**Cut-off level = 5,000**
диалектическая логика (dialectical logic)	фразовое ударение (phrasal stress)
барионный заряд (baryon charge)	блуждающие огни (will-o’-the-wisp)
врождённые идеи (innate idea)	золотой телец (golden calf)
гонка вооружений (arms race)	циркуляция крови (blood motion)
адский огонь (hellfire)	кольцевые гонки (circuit race)
Cut-off level = 10,000	**Cut-off level = 30,000**
грудная железа (breast gland)	концептуальное искусство (conceptual art)

критическая теория (critical theory)	институциональный инвестор (institutional investor)
чесменский бой (battle of Chesma)	земские марки (zhemstvo stamps)
автоматический огонь (automatic fire)	агглютинативные языки (agglutinative language)
личное дворянство (personal nobility)	ненасыщенный пар (unsaturated steam)
Cut-off level = 100,000	**Cut-off level = 150,000**
шлиховой анализ (panning)	ноги прохожих (feet of passers-by)
облеченный тон (invested tone)	разделенный экран (divided screen)
дардские народы (dardsky people)	воркутинская улица (Vorkuta street)
трамвайная археология (tram archeology)	осетинская церковь (Ossetian church)
глухой удар (bump)	старый базар (old market)

Table 5: Examples of MWEs at different levels of the global ranking.

6. Conclusion

In this paper, we described an experiment on MWE extraction from corpora. The novelty of the approach lays in the use of two data sources in parallel, a rich set of features, and advanced learning-to-rank methods applied to the task. The proposed approach outperforms traditional association measures and state-of-the-art classification methods. The method is language-independent and employs limited training data. In the future, we plan to apply the method to the extraction of verbal MWEs.

Acknowledgments

Work on problem definition, survey of related work, experimental design, and feature engineering was carried out by Pavel Braslavski and supported by the Russian Foundation for Basic Research grant no. 14-37-50953. Work on method implementation and evaluation was carried out by Elena Tutubalina and supported by the Russian Science Foundation grant no. 15-11-10019.

References

1. Buckley Chris and Voorhees Ellen M. (2004), Retrieval evaluation with incomplete information, In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 25–32.
2. Church Kenneth Ward and Hanks Patrick (1990), Word association norms, mutual information, and lexicography, Computational linguistics, Vol. 16(1), pp. 22–29.
3. Evert Stefan and Krenn Brigitte (2005), Using small random samples for the manual evaluation of statistical association measures, Computer Speech & Language, 19(4), pp. 450–466.
4. Evert Stefan and Krenn Brigitte (2001), Methods for the qualitative evaluation of lexical association measures. In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp. 188–195.

5. Fazly A. and Stevenson S. (2007), Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures, In Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Association for Computational Linguistics, pp. 9-16.
6. Fothergill Richard and Baldwin Timothy (2011), Fleshing it out: A supervised approach to mwe-token and mwe-type classification, In IJCNLP, pp. 911–919.
7. Freund Yoav, Iyer Raj, Schapire Robert E, and Singer Yoram (2003), An efficient boosting algorithm for combining preferences, The Journal of machine learning research, Vol.4, pp. 933–969.
8. Friedman Jerome H. (2001), Greedy function approximation: a gradient boosting machine, Annals of statistics, pp. 1189–1232.
9. Geng Xiubo, Liu Tie-Yan, Qin Tao, and Li Hang (2007), Feature selection for ranking, In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 407–414.
10. Hartmann Silvana, Szarvas György, and Gurevych Iryna (2012), Mining multiword terms from Wikipedia, Semi-Automatic Ontology Development: Processes and Resources, pp. 226–258.
11. Hasan Kazi Saidul and Ng Vincent (2014), Automatic keyphrase extraction: A survey of the state of the art, Proceedings of the Association for Computational Linguistics (ACL), Baltimore, Maryland: Association for Computational Linguistics.
12. Jiang Xin, Hu Yunhua, and Li Hang (2009), A ranking approach to keyphrase extraction. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, pp. 756–757.
13. Karan Mladen, Snajder Jan, and Basic Bojana Dalbelo (2012), Evaluation of classification algorithms and features for collocation extraction in croatian, In LREC, pp. 657–662.
14. Liu Tie-Yan (2009), Learning to rank for information retrieval, Foundations and Trends in Information Retrieval, Vol. 3(3), pp. 225–331.
15. Medelyan Olena, Frank Eibe, and Witten Ian H (2009), Human-competitive tagging using automatic keyphrase extraction, In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Vol. 3, pp. 1318–1327.
16. Pecina Pavel and Schlesinger Pavel (2006), Combining association measures for collocation extraction, In Proceedings of the COLING/ACL on Main conference poster sessions, Association for Computational Linguistics, pp. 651–658.
17. Ramisch Carlos (2015), Evaluation of mwe acquisition. In Multiword Expressions Acquisition, Springer, pp. 105–125.

18. Ramisch Carlos, Villavicencio Aline, and Boitet Christian (2010), Mwetoolkit: a framework for multiword expression identification, In LREC.
19. Vivaldi Jorge, Cabrera-Diego Luis Adrián, Sierra Gerardo, and Pozzi María (2012), Using Wikipedia to validate the terminology found in a corpus of basic textbooks, In LREC, pp. 3820–3827.
20. Wermter Joachim and Hahn Udo (2006), You can’t beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures for collocation and term extraction, In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 785–792.
21. Wu Qiang, Burges Christopher JC, Svore Krysta M, and Jianfeng Gao (2010), Adapting boosting for information retrieval measures, Information Retrieval, Vol. 13(3), 254–270.