Clustering-based Approach to Multiword Expression Extraction and Ranking

Elena Tutubalina
Higher Institute for Information Technology and Information Systems
Kazan Federal University
Kazan, Russia
tutubalinaev@gmail.com

Abstract

We present a domain-independent clustering-based approach for automatic extraction of multiword expressions (MWEs). The method combines statistical information from a general-purpose corpus and texts from Wikipedia articles. We incorporate association measures via dimensions of data points to cluster MWEs and then compute the ranking score for each MWE based on the closest exemplar assigned to a cluster. Evaluation results, achieved for two languages, show that a combination of association measures gives an improvement in the ranking of MWEs compared with simple counts of co-occurrence frequencies and purely statistical measures.

1 Introduction

Extraction of multiword expressions (MWEs) is a challenging and well-known task, aimed at identifying lexical items with idiosyncratic interpretations that can be decomposed into single words (Sag et al., 2002). In this study, we primarily focus on the extraction of two-word expressions in Russian.

A number of lexical association measures and their combinations have been employed in previous studies about extraction of general-purpose collocations and domain-specific terms (Krenn and Evert, 2001; Pearce, 2002; Evert, 2004; Pecina and Schlesinger, 2006; Hoang et al., 2009; Hartmann et al., 2012). Ranked collocations with higher association scores are selected into the n-best list. These simple approaches are limited by the size of corpora and the effect of low frequency on ranking (Krenn and Evert, 2001; Evert and Krenn, 2005; Bouma, 2009). Most studies regard MWE as a classification task and based on supervised methods to predict the class (collocations or non-collocations) to which an MWE candidate relates (Pecina and Schlesinger, 2006; Ramisch, 2015). There is no labeled training set in Russian for these approaches, and data annotation is time-consuming. The task could be seen as a ranking task: ranking model group comparable entities into queries by criteria and constructing a ranking model using training data with exemplars to predict a ranking score. However, there are no formal principles on how to detect comparable MWEs from general-purpose corpora for Russian. Therefore, in this study we focus on clustering semantically similar MWE candidates using association measures, calculated on a general-purpose corpora and Wikipedia.

A particular general-purpose corpus, such as the Russian National Corpus or the British National Corpus, provides only a partial coverage of the modern language. Although association measures have been widely applied, they have a limitation: the computed probabilities may be small in the particular corpus, which gives a lower rank for MWE in the n-best list. To avoid this situation, we incorporate the standard statistical measure, computed from the general-purpose corpus, with Wikipedia, that contains a vast amount of knowledge (e.g., named entities, domain-specific terms, and disambiguation of word senses).

Given a small number of most representative MWEs as exemplars, our primary goal is to identify MWE noun candidates, considering similarity between a candidate and the exemplars, based on association scores in both resources. Our method consists of three steps: (i) extracting bigrams that serve as MWE candidates, adopting Wikipedia arti-

Proceedings of NAACL-HLT 2015, pages 39–43,
Denver, Colorado, May 31 – June 5, 2015.

cles, and using predefined morphosyntactic patterns; (ii) grouping the candidates using clustering techniques; and (iii) ranking MWE candidates by a score, which is computed based on the distance between the candidate and the closest exemplar multiplied by the percent of exemplars in the cluster. The third step relies on the intuition that MWEs are highly ranked in clusters with a higher number of exemplars due to strong similarity between these expressions.

We demonstrate that combining association measures from two resources is effective, and improvement according to precision-recall curves can be achieved by a small number of measures combined.

2 Related Work

Over the last few decades, a large number of works in computational corpus linguistics have been published concerning the extraction of multiword terms, collocations, and keyphrases that is well described in Evert (2004), Gries, (2013), Hasan and Ng (2014), and Ramisch (2015). The research area covers several different methods, for example, ranking MWEs by association measures (Krenn and Evert, 2001; Pearce, 2002; Evert, 2004; Braslavski and Sokolov, 2006); contrastive filtering of domain-specific MWEs (Bonin et al., 2010); methods that combine statistic measures to find complex ranking functions, using clustering algorithms and neural networks (Pecina and Schlesinger, 2006; Antoch et al., 2013); machine learning approaches to classify MWEs into predefined categories (Pecina and Schlesinger, 2006; Ramisch, 2015); and Wikipedia-based approaches (Medelyan et al., 2009a; Medelyan et al., 2009b).

Many methods combine the different properties of two or more association measures to find high-ranking collocations with a strong association based on these measures (Church et al., 1991; Pecina and Schlesinger, 2006; Liu et al., 2009). Church et al. (1991) used an association measure constructed from mutual information (MI) and t-score formulae with scaling functions for collocation identification. Pecina and Schlesinger (2006) presented supervised methods based on 82 association measures to define a ranker function. They did not select the "best universal method" for combining association measures because the task depends on many factors, such as language and data, among others. Liu et al. (2009) adopted Wikipedia to compute term relatedness based on a vector of Wikipedia concepts for keyphrase extraction. They selected four measures to group terms of a given document based on the semantic relatedness between them. These measures are cosine similarity, Euclidean distance, pointwise mutual information (PMI), and normalized similarity distance. Antoch et al. (2013) combined association measures considered as binary classifiers using receiver operating characteristic curves. They used a hierarchical clustering algorithm to achieve better results by clustering these measures. The authors observed that high efficiency of combining representatives of the clusters of equivalent association measures depends on a dataset. Jain (2010) proposed that there is no single clustering algorithm that is able to outperform other algorithms across all applications.

3 The Clustering-based Approach for Ranking MWEs

In this section, we describe the proposed clustering-based approach. In contrast to classification methods that predict whether a MWE is a true collocation or not, the goal is to determine which MWE candidates are best statistically similar to a small set of exemplars. Exemplars are MWEs (e.g., from the gold standard set) with a rather high degree of association between the word components. We employ Wiktionary to extract MWE exemplars. We perform the clustering of the extracted MWEs using a k-means algorithm and log-likelihood measure.

The proposed approach is composed of three steps: (i) extracting a list of MWEs from Wikipedia article titles, (ii) computing the log-likelihood of the MWE data given the general-purpose corpus and texts from Wikipedia, and (iii) grouping MWE candidates by the k-means clustering algorithm and then ranking cluster points by measuring the distance from these points to the closest exemplar multiplied by the percent of exemplars in the cluster.

3.1 Selecting MWE Candidates

We selected MWE candidates from Wikipedia article titles due to the following reasons: (i) the Russian sentence structure is very flexible, and extraction of bigrams by the patterns, where words are con-

sidered neighbors (adjacent words), is insufficient; and (ii) Wikipedia article titles have explicit phrase boundaries, marked by human editors in Wikipedia markup (Hartmann et al., 2012). The following filter was applied to all the two-word sequences: the candidates were not allowed to contain punctuation marks except hyphenated expressions, and the candidates were not allowed to contain proper names and common geographic locations. The extracted candidates were then filtered by predefined morphosyntactic patterns (e.g., adjective + noun, noun + noun). The morphosyntactic analyzer Mystem[1] and NLTK library are applied for Russian and English, respectively. We used a list of patterns from Braslavski and Sokolov's (2008) and Manning's papers (1999) for texts in Russian and English, respectively.

3.2 Clustering MWE Candidates and Ranking

The proposed approach assigns MWE candidates to the clusters based on the distribution of statistical measures associated with each candidate in general-purpose corpora. The clustering method we apply is k-means that has been widely used with the Euclidean metric for computing the distance between points and cluster centers (Jain, 2010). As indicated from the results, reported in Section 4 of this paper and recent studies (Evert, 2004; Evert and Krenn; 2005), log-likelihood achieves better results than ranking by other statistical measures, such as t-score and MI. Therefore, we compute log-likelihood as statistical characteristics of MWE candidates, based on two different resources of texts.

In this approach, MWE candidates are represented as points in a two-dimensional space, where each dimension represents by log-likelihood. We make assumption that (i) the distribution over all exemplars is similar to a distribution over all words in the corpus, and (ii) MWEs are independently distributed and probabilities are estimated as frequency ratios, which is similar to the naive Bayes assumption (Baker and McCallum, 1998). MWE candidates are ranked by the following formula, that shows the ranking score of MWE *j* in cluster *cl*:

$$score(mwe = j) = (1 - \frac{\min_{i=1,\ldots,n_{cl}} d(j, gs_i)}{r_{cl}}) * np_{cl} \quad (1)$$

[1]Mystem is available at https://tech.yandex.ru/mystem/.

where n_{cl} indicates the number of exemplars in cluster cl, np_{cl} denotes n_{cl} in percent, $d(j, gs_i)$ denotes Euclidean distance between MWE *j* and the exemplar gs_i in cluster cl, r_{cl} denotes radius of cluster cl.

4 Evaluation

We use the Russian National Corpus (RNC) and the British National Corpus (BNC) as the general-purpose corpus of the Russian language and the English language, respectively. For corpora in Russian, we generated frequency lists of bigrams in singular and plural forms. We adopt n-gram data of English Wikipedia and the BNC, extracted by Lin et al. (2010) and Leech and Rayson (2014). We suppose that all MWE candidates occur at least once in corpora due to frequency thresholds in the lists. Table 1 shows MWEs that are top-ranked by our approach.

Russian MWEs	English MWEs
мировая война (*mirovaya voyna*) 'world war'	world war
советский союз (*sovetskiy soyuz*) 'soviet union'	soviet union
настоящее время (*nastoyashchee vremya*) 'present time'	feature film
чемпионат мира (*chempionat mira*) 'world cup'	binomial name
населенный пункт (*naselennyy punkt*) 'human settlement'	world champion
водные ресурсы (*vodnye resursy*) 'water resources'	popular culture

Table 1: Sample of top-ranked collocations.

We adopt Wiktionary as the gold standard dataset for Russian and English due to use of Russian Wiktionary as a data source for WordNet-like resources. The single-word nouns from Wiktionary were used as "raw materials" for the Yet Another RussNet (YARN) project (Braslavski et al., 2014). Comparison of vocabularies in the English and Russian editions of Wiktionary is described in (Krizhanovsky and Smirnov, 2013). The gold standard set for Russian was filtered to remove non-collocations. Table 2 shows a summary of MWEs for two languages.

We compute the precision-recall curves of the n-best lists to evaluate our approach. For comparison, we use n-best lists that are ranked by popular association measures: t-score, log-likelihood, and MI. Wermter and Hahn (2006) proposed that purely association measures could not reveal any significant improvement over co-occurrence frequency. We have also used frequencies of MWEs as a baseline measure for ranking. The types of corpus are followed by a subscript: 1 refers to the general-purpose corpus, and 2 refers to texts from Wikipedia articles.

Language	Russian	English
No. of tokens in the general-purpose corpus	364,881,378	110,691,482
No. of Wikipedia articles	1,172,000+	4,675,000+
No. of MWE candidates	164,805	135,659
No. of MWEs, extracted from Wiktionary	7433	40996
No. of MWEs, selected for the gold standard	3670	40996
Intersection of the sets	2216	6342

Table 2: Summary of the list of MWE candidates.

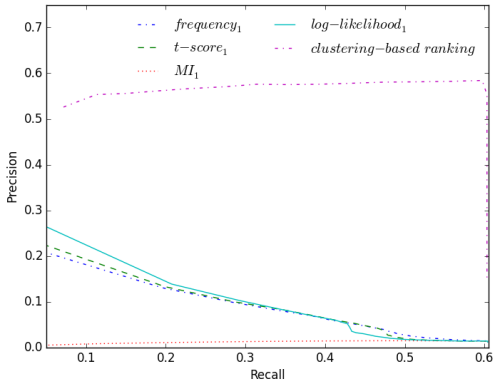

Figure 1: Precision-recall curves of the proposed approach and association measures (for Russian).

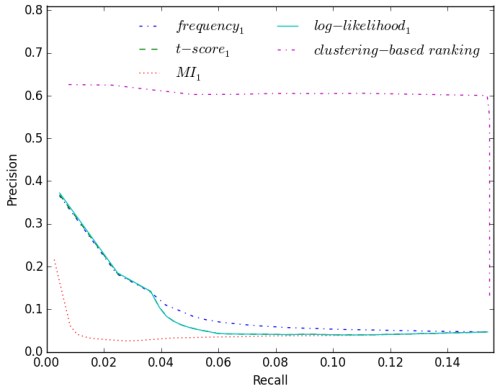

Figure 2: Precision-recall curves of the proposed approach and association measures (for English).

The results, shown in Figures 1 and 2, indicate that the proposed approach outperforms baseline ranking by association measures, but the precision of the n-best list is significantly decreased with increase of recall. In order to evaluate the impact of a varied number of clusters, we conduct experiments on the

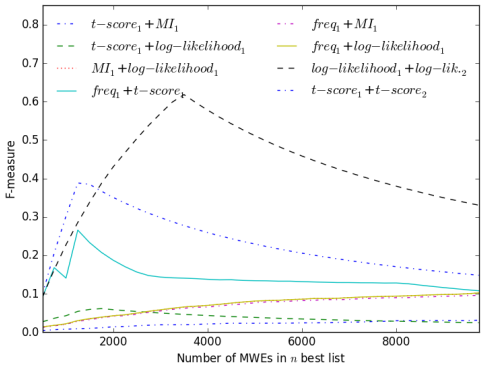

Figure 3: Comparison of F-measure curves of the proposed approach based on different statistical measures.

dataset in Russian using log-likelihood. We change the number of clusters from 5 to 30 to achieve the maximum F-measure with the minimum number of n-best ranked MWEs. Results, shown in Table 3, indicate that n equals 3,500 for each experiment, and the number of clusters is 5.

No. of clusters	P@n	R@n	F-measure
5	0.553	0.6029	0.5769
10	0.5438	0.5928	0.5672
15	0.568	0.5418	0.5546
20	0.5634	0.5374	0.5501
25	0.5431	0.5181	0.5303
30	0.5371	0.5124	0.5245

Table 3: Evaluation results with a varied number of clusters, n equals to 3,500 (for Russian).

To confirm that a combination of association measures from two resources significantly helps in the task of extracting MWEs, we compare our results with different combinations of measures according to F-measure. Figure 3 shows that the combination of log-likelihood, based on two corpora in Russian, gives the best results compared with others.

5 Conclusion and Future Work

In this paper, we proposed a clustering-based approach for the extraction of multiword expressions (MWEs). We incorporated association measures, computed from two corpora, by representing each MWE as a two-dimensional data point. The method assigned MWEs to clusters using k-means clustering and then ranked MWEs by Euclidean distance to the nearest exemplar from the gold standard set. The

efficiency of our approach depends on MWE probabilities in two corpora, and the small set of multiword exemplars is required. For future works, we plan to split MWE candidates into small queries of comparable MWEs by linguistic criteria and then use query-dependent ranking for each query-MWE pair.

Acknowledgments

This work was partially supported by Russian Foundation for Basic Research (Project № 13-07-00773).

References

Antoch J., Prchal L., and Sarda P. 2013. *Combining Association Measures for Collocation Extraction Using Clustering of Receiver Operating Characteristic Curves*. Journal of classification, 30(1):100-123.

Baker L. D. and McCallum A. K. 1998. *Distributional clustering of words for text classification*. Proceedings of the ACM SIGIR conference on Research and development in information retrieval, pp. 96-103.

Bonin F., Dell'Orletta F., Venturi G., & Montemagni S. 2010. *Contrastive filtering of domain-specific multiword terms from different types of corpora*. Proceedings of 23rd International Conference on Computational Linguistics, p. 77

Bouma G. 2009. *Normalized (pointwise) mutual information in collocation extraction*. Proceedings of GSCL, pp. 31-40

Braslavski P. and Sokolov E. 2008. *Comparison of five methods for variable length term extraction*. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog", №. 7, pp. 14. (In Russian)

Braslavski P., Ustalov D., and Mukhin M. 2014. *A spinning wheel for YARN: user interface for a crowdsourced thesaurus*. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 101-104.

Evert S. and Krenn B. 2001. *Multiword expressions: A pain in the neck for NLP*. Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, pp. 188-195

Evert S. 2004. *The statistics of word cooccurrences: word pairs and collocations*.

Evert S. and Krenn B. 2005. *Using small random samples for the manual evaluation of statistical association measures*. Computer Speech & Language, № 4.

Gries S. T. 2013. *50-something years of work on collocations: what is or should be next*. International Journal of Corpus Linguistics, 18(1):137-166.

Gusfield D. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, UK.

Pearce D. 2002. *A Comparative Evaluation of Collocation Extraction Techniques*. LREC.

Pecina P. and Schlesinger P. 2006. *Combining association measures for collocation extraction*. Proceedings of the COLING/ACL on Main conference poster sessions, pp. 651-658.

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).

Hartmann S., Szarvas G., & Gurevych I. 2012. *Mining multiword terms from Wikipedia. Semi-Automatic Ontology Development: Processes and Resources*. Semi-Automatic Ontology Development: Processes and Resources, pp. 226-258.

Hoang H. H., Kim S. N., & Kan M.-Y. 2009. *A re-examination of lexical association measures*. Proceedings of the ACL 2009 Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, pp. 31–39.

Jain A. K. 2010. *Data clustering: 50 years beyond K-means*. Pattern recognition letters, 31(8):651-666.

Lin, D., Church, K. W., Ji, H., Sekine, S., Yarowsky, D., Bergsma, S., & Narsale, S. 2010. *New Tools for Web-Scale N-grams*. LREC.

Krizhanovsky A. A. and Smirnov A. V. 2013. *An approach to automated construction of a general-purpose lexical ontology based on Wiktionary*. Journal of Computer and Systems Sciences International, 52(2):215-225.

Leech G. & Rayson P. 2014. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Manning C. D. 1999. *Foundations of statistical natural language processing*. MIT press.

Medelyan Olena, Frank Eibe, and Witten Ian H. 2009a. *Human-competitive tagging using automatic keyphrase extraction*. Proceedings of the 2009 Conference on EMNLP, pp. 1318–1327.

Medelyan Olena, Milne David, Legg Catherine, and Witten Ian H. 2009b. *Mining meaning from Wikipedia*. International Journal of Human-Computer Studies, № 9 (2009), pp. 716-754.

Ramisch, C. 2015. *Evaluation of MWE Acquisition*. Multiword Expressions Acquisition, pp. 105-125.

Sag Ivan A., Baldwin Timothy, Bond Francis, Copestake Ann, and Flickinger Dan 2002. *Multiword expressions: A pain in the neck for NLP*. Computational Linguistics and Intelligent Text Processing, pp. 38-43

Wermter J. and Hahn U. 2006. *You can't beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures for collocation and term extraction*. The 21st International Conference on Computational Linguistics, pp. 785-792