Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2019”

Moscow, May 29—June 1, 2019

MEASURE CLUSTERING APPROACH TO MWE EXTRACTION[1]

Rossyaykin P. O. (petrrossyaykin@gmail.com),
Loukachevitch N. V. (louk_nat@mail.ru)
Lomonosov Moscow State University, Moscow, Russia

In this paper we present an unsupervised and resource-independent approach to the well-known task of discovery of multiword expressions (MWE) in text corpora. We experimented on extracting Russian nominal phrases (Adj-N and N-N.Gen) relevant for lexical resources (thesauri, WordNet, etc.). Our approach is based on the assumption that idiosyncrasy of MWEs can be due to different properties (morphosyntactic, semantic, pragmatic and statistical), and thus, different types of measures (statistical, context, distributional) are efficient at extracting different MWEs. We propose new context measures as well as an unsupervised method of combining measures in which we cluster vectors of ranks assigned by individual measures. The proposed method accounts for different properties of MWEs and allows surpassing both individual measures and their simple sum/product.

Key words: multiword expressions (MWEs), MWE extraction, association measures, context measures, distributional semantics, clustering, thesaurus, Russian language

ИЗВЛЕЧЕНИЕ MWE НА ОСНОВЕ КЛАСТЕРИЗАЦИИ МЕР

Россяйкин П. О. (petrrosyaykin@gmail.com),
Лукашевич Н. В. (louk_nat@mail.ru)
МГУ им. М. В. Ломоносова, Москва, Россия

Ключевые слова: устойчивые словосочетания, извлечение устойчивых словосочетаний, меры ассоциации, контекстные меры, дистрибутивная семантика, кластеризация, тезаурус, русский язык

[1] The reported study was partially funded by RFBR according to the research project № 18-00-01226 (18-00-01240).

1. Introduction

MWEs, also called collocations or multiword units (MWU), have a long history in NLP. Numerous definitions of MWE were proposed in the literature on both theoretical and computational linguistics. All of them emphasize two core features of MWEs: 1) they are 'words with spaces', i.e. sequences of graphical words not shorter than 2 words and 2) they exhibit unusual, unpredictable properties at any level of linguistic analysis. We, thus, adopt a broad definition proposed by [Baldwin & Kim 2010]:

> "Multiword expressions (MWEs) are lexical items that:
> (a) can be decomposed into multiple lexemes; and
> (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity"

This definition accounts for expressions of very different nature: idioms (*to kick the bucket*), which are semantically idiosyncratic, lexicalized expressions (*black and white television*) displaying statistical idiosyncrasy, terms (*vowel harmony*), which are usually idiosyncratic both semantically and statistically, proper names (*South Korea*), morphologically/syntactically rigid expressions (*by and large*), etc. Information on MWEs of all of these types is crucial for performance in many NLP tasks and applications: machine translation [Ren et al. 2009]; [Carpuat & Diab 2010], syntactic parsing [Korkontzelos & Manandhar 2010], word sense disambiguation [McCarthy et al. 2004], etc.

The number of MWEs in the language is comparable to that of single lexemes or even surpasses it [Jackendoff 1997]; [Sag et al. 2002]. Moreover, new MWEs appear constantly which makes manual compilation of MWE lists inefficient. This prompts the task of automatic MWE extraction (or discovery), which consists in providing a ranked list of expressions that can be either examined and refined by human experts or used in further applications as is.

The goal of this work is to elaborate a method to supplement lexical resources with MWEs. Hence, we focus on nominal phrases (Adj-N and N-N.Gen) of all semantic types mentioned. Despite being composed of multiple lexemes, they should be included in lexical resources as single entries due to 1) their correspondence to single entities on the ontological level and 2) impossibility to account for them with any regular rules of syntax and/or semantics.

We deal with Russian data; however, the methods discussed and proposed in this paper are mostly language-independent — they require only a text corpus and basic pre-processing (lemmatizing and PoS-tagging).

The structure of paper is as follows: we provide a short overview of previous work (paragraph 2), describe our corpus and the set of candidate expressions (3), introduce individual features used in our combinational method: most common lexical association measures, original context measures which yield good results on their own and two state-of-the-art distributional measures (4), introduce a new clustering-based approach to combining measures (5), provide and discuss results (6), (7).

2. Related work

Starting with the papers by [Choueka 1988] and [Church & Hanks 1990] statistical methods prevail in MWE extraction. The only information used by statistical association measures is frequency distribution of words in the corpus, in most cases number of occurrences and co-occurrences of MWEs' components (collocates). Numerous association measures, previously used in other tasks, were adapted to MWE extraction: PMI [Church & Hanks 1990], t-score [Church et al. 1991], log-likelihood ratio [Dunning 1993] are among the most popular.

Crucially, association measures are restricted by design in their ability to discover MWEs. They take advantage of just one property of MWEs (statistical idiosyncrasy), which is irrelevant for certain expressions (e.g. *red tape* with both components being very frequent independently). Moreover, they do not take into account semantic and statistical asymmetry of most MWEs and can be biased to either rare or frequent bigrams (for further criticism of association measures see [Evert 2007: 7.1]).

Alternative approach is based on detecting non-compositionality with the help of either context measures [Nakagawa & Mori 2003], [Riedl & Biemann 2015] or distributional semantics [Lin 1998], [Padó & Lapata 2007], [Van de Cruys & Moirón 2007], [Baroni & Zamparelli 2010]. The introduction of word2vec [Mikolov et al. 2013] triggered a new surge of research in distributional semantics with word embeddings being adapted to different tasks including MWE extraction. In most cases these methods are designed for either particular syntactic patterns (phrasal verbs — [Baldwin et al. 2003], [Salehi et al. 2015]; V-N idioms — [McCarthy et al. 2007], [Senaldi et al. 2016] or lexical types of MWEs [Rodríguez-Fernández et al. 2016], [Enikeeva & Mitrofanova 2017].

Most recent papers deal with combining different methods rather than individual association or distributional measures. [Pecina & Schlesinger 2006] used hierarchical clustering to select a set of statistical and context features and different machine learning algorithms to provide ranking function. [Tsvetkov & Wintner 2011], [Buljan & Šnajder 2017] connected statistical and morphosyntactic measures in a Bayesian network. [Tutubalina & Braslavski 2016] adopted learning-to-rank methods from information retrieval.

Unsupervised approaches to combining measures are much less common. [Zakharov 2017] combined association measures by averaging ranks of MWEs obtained with the use of individual measures. [Tutubalina 2015] used clustering in 2-dimensional space with log-likelihood ratios calculated on 2 different corpora. In contrast to this method, we used one corpus and measures of different nature (statistical, context, distributional) as dimensions. We assume that such an approach allows separating MWEs of different types from free phrases.

3. Data

The corpus we experimented on was composed of news' texts from the Russian Internet published in 2011. We deleted all punctuation, lemmatized and uppercased it, PoS-tagging was used in order to obtain the initial list of candidate Adj-N and N-N bigrams. Bigrams with the observed frequency of less than 200 were excluded, resulting in the list of 37,767 candidate expressions. Given PoS-filtering and a high

frequency threshold, we suppose that our dataset contained no bigrams which systemically were not actual syntactic constituents.

We used the Russian language thesaurus RuThes [Loukachevitch et al. 2014] as our gold standard. Expressions present in it were regarded as actual MWEs (9,837 in total). Our task, thus, was to provide a method which would rank these 9,837 expressions on the top of the list.

4. Individual measures

4.1. Overview

We calculated 22 statistical association measures including the most popular ones (PMI and its variants, t-score, LLR, Dice coefficient, etc.) as well as less common measures which showed good performance in previous comparative studies [Pecina 2008], [Hoang et. al 2009]. 5 asymmetric variants of MI and PMI proposed by [Hoang et al. 2009] and [Carlini 2014] were also added to our comparison.

8 context measures, 4 of which are introduced in this study (see detailed description below), were calculated to obtain a more semantics-based view on our dataset. Formulae for all 30 individual measures are presented in **Table 1**.

Table 1. Statistical association and context measures used for ranking MWE candidates

Name	Formula
frequency	$f(xy)$
PMI	$\log \frac{P(xy)}{P(x)P(y)}$
Sørensen–Dice coefficient (DC)	$\frac{2 * f(xy)}{f(x) + f(y)}$
log-likelihood ratio	$2 \sum_{x,y} f(xy) * \log \frac{p(xy)}{p(x) * p(y)}$
chi-square	$\frac{(f(xy) - \frac{f(x) * f(y)}{N})^2}{f(x) * f(y)}$
Piatetsky-Shapiro coefficient	$P(xy) - P(x) * P(y)$
t-score	$\frac{P(xy) - P(x) * P(y)}{\sqrt{\frac{P(xy)}{N}}}$
geometric mean	$\frac{f(xy)}{\sqrt{f(x) * f(y)}}$
normalized PMI	$\frac{PMI(xy)}{-\log(P(xy))}$

Name	Formula
odds ratio	$\log\frac{(f(xy)+\frac{1}{2})(f(\bar{x}\bar{y})+\frac{1}{2})}{(f(x\bar{y})+\frac{1}{2})(f(\bar{x}y)+\frac{1}{2})}$
Poisson significance measure	$((P(xy)*N-f(xy)*\log(P(xy)*N)+\log(f(xy)!))/\log N$
modified DC	$\log\big(f(xy)\big)*DC(xy)$
Confidence	$\max(P(y\|x),P(x\|y))$
local PMI	$f(xy)*PMI(xy)$
augmented PMI	$\log\frac{P(xy)}{P(x\bar{y})P(\bar{x}y)}$
cubic PMI	$\log\frac{P(xy)^3}{P(x)P(y)}$
normalized MI	$\frac{\sum_{x,y}P(xy)*\log\frac{P(xy)}{P(x)*P(y)}}{-\sum_{x,y}P(xy)*\log P(xy)}$
MI/NF(0.5)	$\frac{MI}{0.5*P(x)+0.5*P(y)}$
PMI/NF(0.77)	$\frac{PMI}{0.77*P(x)+0.23*P(y)}$
MI/NFmax	$\frac{MI}{\max(P(x),P(y))}$
PMI/NFmax	$\frac{PMI}{\max(P(x),P(y))}$
NPMIC	$\frac{PMI(xy)}{-\log(P(x))}$
gravity count (GC)	$\log\frac{f(xy)*\|r(x)\|}{f(x)}+\log\frac{f(xy)*\|l(y)\|}{f(y)}$
modified GC	$\log(\frac{f(xy)*\|r(x)\|}{f(x)}+\frac{f(xy)*\|l(y)\|}{f(y)})$
type-LR	$\sqrt{\|r(x)\|*\|l(y)\|}$
type-FLR	$\frac{f(xy)}{typeLR(xy)}$
context intersection (CI)	$\frac{\|l(xy)\cap l(x)\|}{\|l(x)\|}*\frac{\|r(y)\cap r(xy)\|}{\|r(y)\|}$
independent CI	$\frac{\|l(xy)\cap l(xW)\|}{\|l(xW)\|}*\frac{\|r(Wy)\cap r(xy)\|}{\|r(Wy)\|}$
CI*freq	$f(xy)*CI(xy)$
ICI*log(freq)	$\log f(xy)*ICI(xy)$

Where N is the number of tokens in corpus, xy — bigram consisting of words x and y, $f(x)$ is the observed frequency of the word x, $P(x) = f(x)/N$, $\bar{x}$ stands for any word except x, $r(x)$ is a set of unique words occurring in corpus immediately to the right from the word x, $l(x)$ is a set of unique words which occur in corpus immediately to the left from the word x, $\sum_{x,y} A(x,y) = A(x,y) + A(x,\bar{y}) + A(\bar{x},y) + A(\bar{x},\bar{y})$, W stands for any word which does not form a candidate expression with an adjacent word (x in xW or y in Wy)

4.2. Context measures

In their work on automatic term recognition [Nakagawa & Mori 2003] proposed to calculate how many distinct compound nouns contain the simple noun in question as their part in a given corpus, i.e. to build sets of unique words which occur immediately to the left and to the right from the word W. Cardinalities of these sets are multiplied in the scoring function. We use this idea to model lexical rigidity (non-substitutability) of MWE components. In our measure type-LR (see **Table 1**) we take geometric mean of the number of unique words which can occur in the first and in the second position of the MWE under consideration (with the other word being fixed). Our assumption is as follows: the fewer words occur to substitute components of the given bigram, the higher its probability to be an actual MWE.

Taking into account that usual statistical association measures and context measures use different properties of MWEs, we also incorporated the observed frequency into type-LR. This modification (type-FLR) gave a significant increase in average precision (see paragraph 6 for results).

The other group of measures proposed is based on the idea of [Riedl & Biemann 2015] that MWEs tend to have single-word synonyms and, thus, contexts of MWEs of different lengths are similar to that of single words. We took another perspective on the context data comparing immediate contexts of MWEs with those of their components (see context intersection (CI) and independent context intersection (ICI) in **Table 1**). We also combined CI and ICI with either raw observed frequency or its binary logarithm, the best combinations are present in **Table 1**.

Using context and mixed measures introduced in this paragraph we achieved average precision comparable to that of the best measures included in comparison. ICI multiplied by logarithm of frequency significantly surpassed all other individual measures (see **Table 2** below).

4.3. Distributional measures

The only distributional measure we used is DFsing/DFthes proposed in [Loukachevitch & Parkhomenko 2018]. It showed extremely high average precision on the same data. We used the model trained by the authors of the original paper with word2vec [Mikolov et al. 2013] and the following parameters: vector size 200, window size 3, min_count 3 (other parameters left default). Bigrams from dataset were concatenated into single tokens using underscores ('x y' -> 'x_y') to make word-2vec able to build vectors for them. DFsing is calculated as the similarity between the phrase vector $v(xy)$ and vector of the most similar single word w; the word should be different from the phrase components.

$DFsing = \max(\cos(v(xy), v(w)))$, where w is a word from the model vocabulary distinct from x and y.

Given the task of thesauri extension [Loukachevitch & Parkhomenko 2018] also proposed modification of DFsing, which calculates the maximal cosine similarity of a phrase with the existing text entries of the basic thesaurus (RuThes in our case) and orders phrase candidates according to decreasing value of similarity with thesaurus entries (single words or phrases).

$DFthes = \max(\cos(v(xy), v(te)))$, where te is a thesaurus entry (word or phrase).

Note that it is the only measure requiring external resources and its inclusion is not crucial for our combinational method. Following [Loukachevitch & Parkhomenko 2018] we also multiplied DFsing and DFthes by binary logarithm of frequency (see **Table 2** below).

5. Clustering

The idea behind combining measures is the following: an MWE can be indistinguishable from free phrases according to, for example, its frequency properties but stick out as for context or distributional properties (or vice versa). The example of simple combination of two features is provided at **Figure 1**. Values of PMI3 and type-FLR (normalized with binary logarithm), which do not use any common data when calculated, serve as coordinates in 2-dimensional space. It is clearly visible that MWEs (red dots) and free phrases (blue dots) tend to cluster. However, if we look at the border zone of the clusters we will see that expressions of two classes are still substantially mixed and there is no hyperplane which would be able to separate them with high precision.

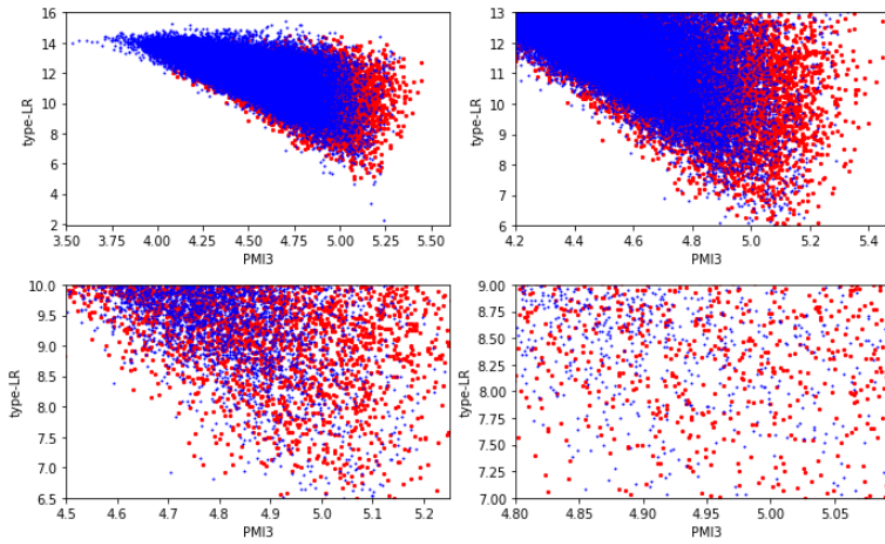

Figure 1. Distribution of expressions in 2-dimensional space with binary logarithms of values assigned by PMI³ and type-LR used as coordinates

This prompts us to increase the number of features, i.e. dimensionality of feature space. The question is how to map feature vectors to numbers (for ranking purposes) without training any classifier. Table 3 below shows that simple sum or product of coordinates (= feature values) tends to yield deteriorating results with the increasing number of features.

The alternative approach proposed by us aims at preserving dimensionality rather than simply compressing vectors into numbers. First of all, since values assigned by different measures vary considerably, for clustering purposes they were mapped to ranks with binary logarithms[2] taken to make contribution of higher ranks more significant. As a result, every candidate expression was associated with a vector consisting of logarithms of ranks assigned by individual features. These vectors were divided into 2 clusters using implementations of k-means and agglomerative hierarchical clustering (with Ward linkage strategy) in scikit-learn library of Python. We assume that the smaller cluster corresponds to actual MWEs with the larger one consisting mostly of free phrases.

Since we are interested in ranking, rather just classifying, bigrams, we used the following centroid-oriented scoring function:

$$rank(xy) = d(xy, \vec{\mu}_0) - d(xy, \vec{\mu}_1)$$

Where $d(a,b)$ is Euclidean distance, xy is the vector of an expression 'x y', $\vec{\mu}_0$ is the centroid of the larger cluster, and $\vec{\mu}_1$ is the centroid of the smaller cluster.

6. Results

To evaluate the list rankings, we utilized uninterpolated average precision measure (AP), which achieves the maximal value (1) if all expressions of the positive class are located in the beginning of a list without any interruptions. AP at the level of k first candidates is calculated as follows:

$$AP@k = \frac{1}{m}\sum_{i=1}^{k}(r_i * (\frac{1}{i}\sum_{1\le j\le i} r_j))$$

Where $r_i = 1$ if i-th candidate belongs to the positive class, $r_i = 0$ otherwise, m is the number of elements in the positive class.

Table 2 shows the results for individual measures with the best ones being compared at Figure 2.

When combining measures we have experimented on the set of 8 measures with the highest AP. They include measures of all three types — statistical (cubic PMI and LLR), context (type-FLR and ICI*log(freq)) and distributional (DFsing, DFthes and variants multiplied by binary logarithms of frequency). We performed 2 variants of clustering (k-means and agglomerative) with the scoring function defined in paragraph 5 on all 247 feature subsets with more than 1 element. For every feature subset we also tried to combine logarithms of ranks by simply multiplying or summing them

[2] Sigmoid function can be used instead as was proposed by an anonymous reviewer.

up. Table 3 shows the best results, all of them except for the first one were obtained using agglomerative clustering.

Table 2. Average precision of individual measures

measure	AP@100	AP@500	AP@1000	AP@2500
Statistical association measures (except PMI and MI variants)				
frequency	0.725	0.734	0.698	0.615
PMI	0.518	0.544	0.545	0.532
Sørensen–Dice coefficient	0.697	0.683	0.674	0.636
LLR	0.778	**0.802**	**0.780**	**0.705**
chi-square	0.699	0.704	0.693	0.657
Piatetsky-Shapiro	0.726	0.740	0.706	0.625
t-score	0.727	0.743	0.710	0.631
geometric mean	0.699	0.704	0.693	0.657
odds ratio	0.559	0.598	0.59	0.562
Poisson	0.775	0.799	0.777	0.702
modified DC	**0.827**	0.736	0.713	0.664
confidence	0.56	0.647	0.642	0.608
Symmetric variants of PMI and MI				
local PMI	0.768	0.792	0.768	0.693
augmented PMI	0.567	0.6	0.591	0.562
cubic PMI	**0.907**	**0.821**	**0.795**	**0.726**
NPMI	0.653	0.663	0.651	0.615
NMI	0.687	0.672	0.666	0.63
Asymmetric variants of PMI and MI				
MI / NF(0,5)	0.64	**0.655**	**0.652**	**0.618**
PMI / NF(0,77)	0.508	0.5	0.495	0.471
MI / NFmax	**0.641**	0.646	0.643	0.609
PMI / NFmax	0.452	0.476	0.472	0.447
$NPMI_c$	0.567	0.529	0.516	0.497
Context measures				
gravity count	0.708	0.694	0.662	0.594
modified GC	0.703	0.695	0.666	0.599
type-LR	0.521	0.563	0.553	0.529
type-FLR	0.818	0.825	0.796	0.74
CI	0.748	0.789	0.783	0.743
ICI	0.894	0.869	0.843	0.78
CI*freq	0.902	0.866	0.847	0.777
ICI*log(freq)	**0.915**	**0.879**	**0.855**	**0.789**
Distributional measures				
DFsing	0.846	0.770	0.694	0.583
DFsing*log(freq)	0.929	0.877	0.834	0.731
DFthes	**0.953**	0.853	0.823	0.759
DFthes*log(freq)	0.950	**0.910**	**0.879**	**0.807**

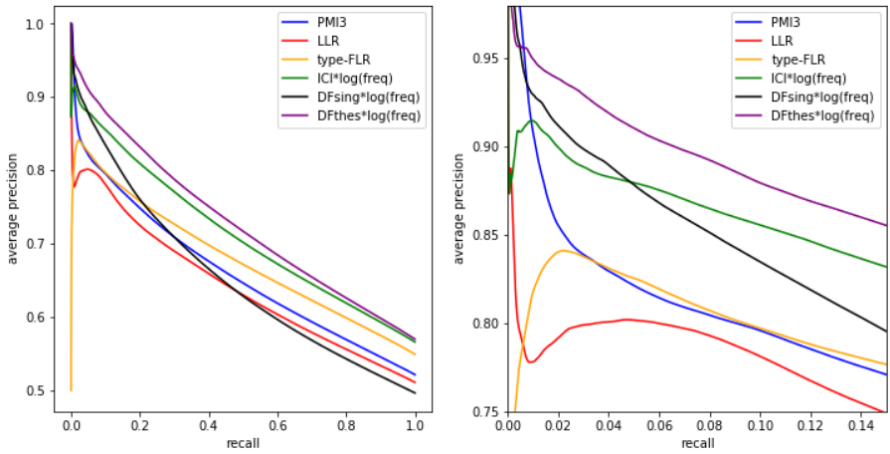

Figure 2. Average precision of the best individual measures with the recall up to 1 and 0.15

Table 3. Best variants of measures' combinations

Method	measures used	AP@100	AP@500	AP@1000	AP@2500
sum of ranks' logarithms	type-FLR, DFthes, DFsing*log(f)	0.976	0.945	**0.92**	**0.847**
agglomerative clustering of ranks' logarithms	type-FLR, ICI*log(f), DFthes, DFsing*log(f)	0.986	0.94	0.907	0.84
	LLR, type-FLR, DFsing, DFthes, DFthes*log(f)	**0.991**	**0.955**	0.917	**0.847**
	LLR, type-FLR, ICI*log(f), DFsing, DFthes, DFsing*log(f), DFthes*log(f)	0.988	0.95	0.914	0.844

7. Discussion

Although simple sum of ranks provides more consistent results, especially when using low amounts of features, the best results are achieved with clustering-based ranking function applied to larger subsets of features. Importantly, two best variants use measures of all three types. Note also that except cubic PMI all of the measures we tried to combine appear in the best setups at least twice.

It is well-known that statistical association measures tend to be biased to either rare or frequent expressions [Evert 2007], [Bouma 2009]. Use of context and distributional measures allows promoting 'unusual' expressions. Table 4 shows top-10 lists obtained with type-FLR and DFthes which turned out to be the most robust features for combining purposes appearing in all four best combinational setups (see Table 3). Finally, Table 5 shows top-20 list obtained with the best variant of clustering (LLR, type-FLR, DFsing, DFthes, DFthes*log(freq)). Note that there are no common expressions in these three lists.

Table 4. Top-10 bigrams extracted with type-FLR and DFthes

type-FLR		DFthes	
Едиот Ахронот 'Yedioth Ahronoth'	N	детский сад 'kindergarten'	T
заработная плата 'salary'	T	Европейский союз 'European Union'	T
правоохранительный орган 'law enforcement agency'	T	атомная электростанция 'nuclear power station'	T
Централ Партнершип 'Central Partnership'	N	атомная станция 'nuclear station'	T
точка зрения 'point of view'	T	международное сообщество 'international community'	T
рубрика Автоновости 'Autonews column'	N	мировое сообщество 'world community'	T
уголовное дело 'criminal case'	T	генеральная прокуратура 'prosecutor-general's office'	T
Ближний Восток 'Near East'	T	районный суд 'district court'	T
алкогольное опьянение 'alcohol intoxication'	T	государственный бюджет 'government budget'	T
тройская унция 'Troy ounce'	T	следственный изолятор 'detention center'	T

Table 5. Top-20 bigrams extracted with the best combinational ranking function

1–10		11–20	
ремонтные работы 'reconditioning'	T	Донецкая область 'Donetsk region'	T
Красноярский край 'Krasnoyarsk region'	T	Оренбургская область 'Orenburg region'	T
исполнительный директор 'executive director'	T	антиправительственное выступление 'antigovernment rally'	T
административная ответственность 'administrative liability'	T	избирательная кампания 'election campaign'	T
товарищеский матч 'exhibition game'	T	киевское Динамо 'Kievan Dynamo'	T
Приморский край 'Primorsky kray'	T	наркотическое средство 'narcotic substance'	T
Томская область 'Tomsk region'	T	добыча нефти 'oil extraction'	T
мобильный телефон 'mobile phone'	T	силовая структура 'uniformed service'	T
сотовый оператор 'mobile network operator'	T	общеобразовательная школа 'comprehensive school'	T
федеральный бюджет 'federal budget'	T	Иркутская область 'Irkutsk region'	T

8. Conclusion

In this paper we have introduced simple yet highly efficient unsupervised approach to extracting MWEs appropriate for lexical resources. We have shown that the choice of features is crucial for the efficiency of combinational MWE extraction. Comparing 22 statistical association measures, 8 context measures (4 of which were introduced in this paper) and 4 distributional measures we showed that ranked lists extracted by measures of different types exhibit significant variation. We also showed that the highest average precision is achieved with the help of measures which utilize both frequency and context/distributional information.

We tried out two unsupervised approaches to combining measures: simple sum or product and clustering. Dividing feature vectors of MWEs into 2 clusters and computing distances to their centroids allows incorporating larger number of measures with higher average precision. We leave for further research testing the stability of average precision given particular subset of measures and varying input data.

References

1. *Baldwin T., Bannard C., Tanaka T., Widdows D.* (2003), An empirical model of multiword expression decomposability. In Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, pages 89–96.
2. *Baldwin T., Kim S. N.* (2010), Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, Handbook of Natural Language Processing, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2 edition.
3. *Baroni M., Zamparelli R.* (2010), Nouns are vectors, adjectives are matrices: Representing adjective–noun constructions in semantic space. In Proceedings of the EMNLP-2010, pp. 1183–1193.
4. *Bouma G.* (2009), Normalized (pointwise) mutual information in collocation extraction. In From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009, volume Normalized, pages 31–40, Tübingen.
5. *Buljan, Šnajder J.* (2017), Combining Linguistic Features for the Detection of Croatian Multiword Expressions. Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), Valencia, 194–199.
6. *Carlini, R., Codina-Filbà, J., Wanner, L.* (2014), Improving collocation correction by ranking suggestions using linguistic knowledge. Proceedings of the 3rd Workshop on NLP for computer-assisted language learning, Uppsala, Sweden.
7. *Carpuat M., Diab M.* (2010), Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 242–245. Association for Computational Linguistics.
8. *Choueka Y.* (1988), Looking for needles in a haystack. In Proceedings of RIAO '88, pages 609–623.

9. *Church K., Hanks P.* (1990), Word Association Norms, Mutual Information, and Lexicography. In Proceedings of ACL, pages 76–83, 1989.
10. *Church K., Gale W. A., Hanks P., Hindle D.* (1991), Using statistics in lexical analysis. In Lexical Acquisition: Using On-line Resources to Build a Lexicon, pages 115–164. Lawrence Erlbaum.
11. *Van de Cruys, T., Moirón, B. V.* (2007), Semantics-based multiword expressions extraction. In: Proceedings of the Workshop on A Broader Perspective on Multiword Expression, pp. 25–32.
12. *Dunning T. E.* (1993), Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), 61–74.
13. *Enikeeva E. V., Mitrofanova O. A.* (2017), Russian Collocation Extraction Based on Word Embeddings. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017", Vol. 1, 2017, pp. 52–65.
14. *Evert S.* (2007), Corpora and collocations. Extended Manuscript of Chapter 58 of A. Lüdeling and M. Kytö, 2008, Corpus Linguistics. An International Handbook, Mouton de Gruyter, Berlin.
15. *Hoang H. H., Kim S. N., Kann M.* (2009), A re-examination of lexical association measures. In Proceedings of the ACL 2009 Workshop on MWEs, pages 31–39, Singapore.
16. *Jackendoff R.* (1997), The Architecture of the Language Faculty. Number 28 in Linguistic Inquiry Monographs. MIT Press, Cambridge, MA, USA. 262 p.
17. *Korkontzelos I., Manandhar S.* (2010). Can recognising multiword expressions improve shallow parsing? In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 636–644. Association for Computational Linguistics.
18. *Lin D.* (1998), Automatic retrieval and clustering of similar words. In Proceedings of COLING/ACL-98, pages 768–744, Montreal.
19. *Loukachevitch N., Dobrov B., Chetviorkin I.* (2014). "Ruthes-lite, a publicly available version of thesaurus of russian language ruthes." Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", Bekasovo, Russia. 2014.
20. *Loukachevitch N., Parkhomenko E* (2018), Recognition of multiword expressions using word embeddings // Artificial Intelligence. RCAI 2018. — Vol. 934 of Communications in Computer and Information Science. — Springer Cham, 2018. — P. 112–124.
21. *McCarthy D., Koeling R., Weeds J., Carroll J.* (2004), Finding predominant word senses in untagged text. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pp. 280–287.
22. *Mikolov T., Chen K., Corrado G., Dean J.* (2013), Efficient Estimation of Word Representations in Vector Space. In: Proceedings of Workshop at ICLR.
23. *Nakagawa H., Mori T.* (2003), Automatic Term Recognition based on Statistics of Compound Nouns and their Components // Terminology. — 2003. — Vol. 9, №2. — P. 201–219.

24. *Padó S., Lapata M.* (2007), Dependency-based Construction of Semantic Space Models. Computational Linguistics, 33(2):161–199.
25. *Pecina P., Schlesinger P.* (2006), Combining association measures for collocation extraction. In Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006), Sydney, Australia.
26. *Pecina P.* (2008), A Machine Learning Approach to Multiword Expression Extraction. In Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008), pages 54–57, Marrakech, 2008.
27. *Ren Z., Lü Y., Cao J., Liu Q., Huang Y.* (2009), Improving statistical machine translation using domain bilingual multiword expressions. In Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, pages 47–54. Association for Computational Linguistics.
28. *Riedl M., Biemann C.* (2015), A single word is not enough: Ranking multiword expressions using distributional semantics. In Proceedings of EMNLP 2015, pages 2430–2440, Lisbon.
29. *Rodríguez-Fernández S., Anke L., Carlini R., Wanner L.* (2016), Semantics-driven recognition of collocations using word embeddings'. In: Proceedings of the 2016 Annual Meeting of the Association for Computational Linguistics (ACL), Berlin, Germany.
30. *Sag I. A., Baldwin T., Bond F., Copestake A., Flickinger D.* (2002), Multiword expressions: A pain in the neck for NLP. In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002), pages 1–15, Mexico City, Mexico.
31. *Salehi B., Cook P., Baldwin T.* (2015), A word embedding approach to predicting the compositionality of multiword expressions. In NAACL-HTL, pages 977–983, Denver, Colorado, 2015.
32. *Senaldi M. S. G., Lebani G. E., Lenci A.* (2016), Lexical Variability and Compositionality: Investigating Idiomaticity with Distributional Semantic Models. Proceedings of the 12th Workshop on Multiword Expressions (MWE 2016): 21–31.
33. *Tsvetkov Y., Wintner S.* (2011), Identification of multi-word expressions by combining multiple linguistic information sources. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 836–845. Association for Computational Linguistics.
34. *Tutubalina E.* (2015), Clustering-based Approach to Multiword Expression Extraction and Ranking. In NAACL-HTL, pages 39–43, Denver, Colorado, 2015.
35. *Tutubalina E. V., Braslavski P. I.* (2016), Multiple features for multiword extraction: A learning-to-rank approach// Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016", Vol. 1, 2016, pp. 782–791.
36. *Zakharov V.* (2017), Automatic Collocation Extraction: Association Measures Evaluation and Integration // Computational Linguistics and Intellectual Technologies: Papers from the Annual conference "Dialogue". Volume 1 of 2. Computational Linguistics: Practical Applications. — Moscow: RSUH, 2017. — P. 396–407.