An Experimental Study of Term Extraction for Real Information-Retrieval Thesauri

Natalia Loukachevitch
Lomonosov Moscow State University
louk_nat@mail.ru

Michael Nokel
Lomonosov Moscow State University
mnokel@gmail.com

Abstract

Models for effective term extraction can depend on the type of a terminological resource under construction. In this paper we study term extraction models for real-working information-retrieval thesauri. The first thesaurus is the English version of EuroVoc thesaurus, the second one is the Russian Banking thesaurus. We study single-word and two-word term extraction separately to reveal the best features and feature combinations, compare best models for two thesauri. In particular, we found for this type of terminological resources the use of association measures does not improve the quality of two-word term extraction based on combining multiple features.

1 Introduction

Automatic term extraction from texts of a specific domain is one of the well-studied applications in natural language processing and document analysis. During many years of research a lot of useful features of domain term extraction were proposed, including frequency-based and context-based features, word association measures, etc. ((Daille, 1995), (Zhang, 2008)).

Since these features characterize various properties of terms, machine-learning models based on multiple features are now increasingly used for term extraction. It was shown that such models can work considerably better than those based on single features ((Aze et al., 2005), (Loukachevitch, 2012)). Nevertheless, the significance of particular features for term extraction by machine learning depends on several important aspects concerning the domain in the target text collection, structure of extracted terms, and type of a terminological resource to be developed.

Firstly, specific domains vary in their scope (e.g., the broad social-political domain vs. the relatively narrow banking domain). Besides, domain-specific languages vary in their closeness to the general language (e.g. banking vs. immunology domain). This enhances or diminishes the role of a reference text collection required to calculate some term features (usually, a news collection or a national corpus is used).

Secondly, terms may be single-word and multi-word. To extract single-word terms, word association measures (mutual information, t-score, etc.) are not applicable; extraction of three-word and longer terms requires special forms of association measures. It means that extraction models for terms of different lengths can differ.

At last, terms are extracted for various types of terminological resources: terminological dictionaries, information-retrieval thesauri, ontologies for NLP. Dictionaries are mainly intended to supply terms with definitions, whereas information-retrieval thesauri are to provide concepts (descriptors) for domain-specific applications (Z39.19, 2005).

For example, such terms from EuroVoc information-retrieval thesaurus as *agricultural product, milk product, European party, economic consequence* denote important concepts in the contemporary socio-political life of European Union, however, it is difficult to imagine these terms as entries in terminological dictionaries. Therefore, a particular type of a terminological resource needs specialized term extraction models (Loukachevitch, 2012).

In this paper we consider the term extraction task specially for thesauri intended to be used in the information retrieval context (search, categorization, clustering and other applications), because we suppose that such terminological resources have specific properties partially explained in specialized standards (Z39.19, 2005).

For this task we experimentally study machine-learning models based on various features for term extraction. Our study is based on two manually created thesauri and two languages: the English version of Eurovoc thesaurus and the Russian Banking thesaurus. We restrict our study to single-word and two-word terms to compare the extraction models for the most frequent types of terms.

2 Related Work

Machine-learning or combined approaches to automatic term extraction were studied in a number of works: (Vivaldi et al., 2001), (Aze et al., 2005), (Foo and Merkel, 2010), (Zhang, 2008), (Loukachevitch, 2012).

In most works automatically extracted terms are evaluated on the basis of available terminological resources or expert annotations of domain terms ((Daille, 1995), (Church and Hanks, 1990), (Dunning, 1993), (Church and Gale, 1995)). If to consider evaluation of machine-learning models for term extraction, in (Aze et al., 2005) experiments were fulfilled for texts in biological and human resources domains with expert annotation of domain terms. In the work (Foo and Merkel, 2010) two patent collections with term pre-annotation were studied. (Zhang, 2008) extracted terms from the Genia corpus, for which Genia ontology was created, and also utilized an artificial corpus of Wikipedia articles with expert annotation of terms.

In contrast to the above-mentioned works, in our study of term extraction we focus on the specific type of terminological resources – thesauri intended for information-retrieval applications. We take the well-known terminological resource EuroVoc and Banking thesaurus created for the Central Bank of the Russian Federation. Both resources are used in indexing and retrieval of documents in real information-retrieval systems.

3 Resources

3.1 EuroVoc Thesaurus and Europarl Text Collection

For the English part of our study we took EuroVoc thesaurus and Europarl parallel corpus. EuroVoc is an official thesaurus of the European Union and is intended for manual indexing of EU parliamentary documents. It is a multidisciplinary thesaurus covering the EU activities and containing terms in 22 languages of the EU. The English version of EuroVoc comprises 15161 terms (`http://eurovoc.europa.eu/drupal`).

The Europarl parallel corpus was extracted from the proceedings of the European Parliament (`http://www.statmt.org/europarl/`). The English part includes almost 54 mln. words.

In fact, EuroVoc thesaurus is intended just for the description of Europarl documents. Therefore, we can model how EuroVoc thesaurus could be developed from the Europarl corpus. EuroVoc represents a broad socio-political domain, and its language is close to general English.

3.2 Banking Thesaurus for the Central Bank and Articles from Online Magazines

For the Russian part of our study we took the Banking thesaurus created for the Central Bank of the Russian Federation. It is used in an information-retrieval system for indexing, search and vizualization of information and as a basis for text categorization. The thesaurus includes about 15 thousand terms and comprises the terminology of banking activity, banking regulation, monetary politics and macroeconomics.

As an appropriate text collection we took 10422 Russian articles from various on-line magazines: Auditor, RBC, Banking Magazine, etc. These documents contain almost 15.5 mln. words.

Since the banking thesaurus is used in real information retrieval tasks, we can model how it could be developed from the banking text corpus. In contrast to the broad socio-political domain of EuroVoc, this thesaurus represents relatively narrow banking domain, and its language is not so close to general language.

4 Features for Term Extraction

In our study we investigated single-word and two-word term extraction separately in order to have possibility to compare corresponding extraction models. As single-word term candidates we consider only *Nouns* and *Adjectives* (for Russian language) and *Nouns* (for English language); as two-word candidates we consider only *Adjective + Noun* and *Noun + Noun* (for Russian language) and *Adjective + Noun*, *Noun + Noun*, and *Noun +*

of + *Noun* (for English language) since they cover the majority of terms.

We use several types of enough known features for term extraction proposed in previous works and relatively new topic-based features proposed in (Bolshakova et al., 2013).

4.1 Traditional Features

The first type of traditional features is **frequency-based features**. The main assumption is that terms differ in their frequency and the distribution from other words in the target corpus. We consider the following 8 features: *Term Frequency in the collection (tf), Document Frequency (df), TF-IDF, TF-RIDF* (Church and Gale, 1995), *Domain Consensus* (Sclano and Velardi, 2007), *Term Contribution, Term Variance Quality, Term Variance* (Liu et al., 2005).

The second type of traditional features is based on the **target and reference corpora** and supposes that term frequencies in the target and reference corpora should be significantly different. We consider 9 such features, namely: *Weirdness* (Ahmad et al., 1999), *corpus-based TF-IDF* (where *TF* is taken from the target corpus, and *IDF* is taken from the reference corpus), *Relevance* (Peñas et al., 2001), *Contrastive* (Basili et al., 2001) and *Discriminative* (Wong et al., 2007) *Weights*, *Lexical Cohesion* (Park et al., 2002), *Reference Weight, KF-IDF* (Kurz and Xu, 2002), *Loglikelihood* (Gelbukh et al., 2010). In our study n-gramm statistics from British National Corpus (`http://www.natcorp.ox.ac.uk/`) and Russian National Corpus (`http://www.ruscorpora.ru`) were used as statistical data of a reference corpus for English and Russian collections correspondingly.

The third type of traditional features comprises **word-association measures** estimating mutual correlation of term candidate usage. They are primarily intended for two-word collocation extraction and are not applicable for single-word term extraction. We consider 19 word association measures: *Mutual Information (MI)* (Church and Hanks, 1990), *Augmented MI* (Zhang, 2008), *Cubic MI* (Daille, 1995), *Normalized Pointwise MI* (Bouma, 2009), *True MI, Dice Coefficient (DC)* (Smadja et al., 1996), *Modified DC, Generalized DC* (Park et al., 2002), *T-Score, Z-Score, Symmetric Conditional Probability* (Lopes and Silva, 1999), *Simple Matching Coefficient, Kulczinksy Coefficient, Ochiai Coefficient, Yule Coefficient, Jaccard Coefficient* (Daille, 1995), *Chi Square, Loglikelihood Ratio* (Dunning, 1993), *Gravity Count* (Daudarvičius and Marcinkevičiené, 2005).

The last type of traditional features is **context-based features** that account for phrases encompassing term candidates and their left and/or right context. We define a context of a term candidate as the bounds of encompassing noun phrases. In our study 11 known context-based features were considered: *C-Value, NC-Value* (Frantzi and Ananiadou, 1994), *MNC-Value, Token-LR, Token-FLR, Type-LR, Type-FLR* (Nakagawa and Mori, 2003), *Sum3, Sum10, Sum50, Insideness* (Loukachevitch, 2012).

Besides, we propose a novel context-based feature: *Modified Gravity Count (MGCount)*. It is based on Gravity Count association measure described in (Daudarvičius and Marcinkevičiené, 2005). *MGCount* for xy phrase is calculated as follows:

$$MGCount = \log\left(\frac{f(xy)l(x)}{f(x)} + \frac{f(xy)r(y)}{f(y)}\right) \tag{1}$$

where $f(x)$ is the frequency of x, $f(y)$ is the frequency of y, $f(xy)$ is the frequency of xy phrase, $l(x)$ is the number of different words to the left of x, and $r(y)$ is the number of different words to the right of y; $l(x)$ and $r(y)$ are considered only within the bounds of encompassing noun phrases. Our modification changed internal proportion $\frac{r(x)}{f(x)}$ to external proportion $\frac{l(x)}{f(x)}$ (and the same with the second component of the sum), thus the measure was transformed from the association measure to the context one.

4.2 Topic-Based Features

The next type comprises features based on so-called **topic models** (Blei and Lafferty, 2009). Topic models are intended to describe texts in terms of their topics, they determine, which topics are related to each document, and which words (or phrases) form each topic. In fact, each topic is represented as a list of frequently co-occurring words (or bigrams) ordered by descending degree of belonging to it. As an example, the first five words and bigrams from the top of four randomly selected topics of the English corpus along with their probabilities of belonging are presented in the Table 1.

Topic #1		Topic #2	
Single-word	*Probability*	*Two-word*	*Probability*
Latin	0.021	European union	0.012
America	0.02	Young people	0.005
American	0.012	European council	0.004
United	0.009	United state	0.003
State	0.007	Youth program	0.002
Topic #3		**Topic #4**	
Single-word	*Probability*	*Two-word*	*Probability*
Audiovisual	0.013	Central bank	0.005
Film	0.011	European central	0.003
Television	0.01	Natural resource	0.002
Medium	0.008	Novel food	0.002
Broadcasting	0.006	Monetary policy	0.002

Table 1: Examples of revealed subtopics

Typically, there are two types of topic models: non-probabilistic ones that are based on hard clustering methods (K-Means, hierarchical agglomerative clustering, etc.) and probabilistic ones (PLSI, LDA, etc.) that represent each document as a mixture of topics and each topic is considered as a probabilistic distribution over words (Blei and Lafferty, 2009), (Bolshakova et al., 2013).

The topic-based features are relatively new and are obtained by revealing topics in the target text corpus. These features account for the idea that domain terms should usually correspond to some subtopics of the domain. As it was shown that *NMF (Non-Negative Matrix Factorization)* algorithm with KL-divergence minimization is the best topic model in terms of terminology extraction (Bolshakova et al., 2013), we applied it to reveal subtopics, as well as probabilities in them. Basically, given a non-negative term-document matrix V, this algorithm tries to find non-negative term-topic matrix W and topic-document matrix H, such that $V = WH$. We consider the version of NMF that minimizes Kullback-Leibler divergence $D(V||WH)$ (Lee and Seung, 2000).

We consider the following 7 topic-based features: *Term Frequency, TF-IDF, Domain Consensus, Maximum Term Frequency* (Bolshakova et al., 2013), *Term Score (TS)* (Blei and Lafferty, 2009), *TS-IDF, Maximum Term Score*. Most of these features are extensions of the standard frequency-based features applied to the revealed subtopics, considering probabilities of the term candidates in topics as frequencies (cf. Table 2; $P_i(w)$ denotes a probability of the term candidate w in the topic i, and K is the number of topics).

Feature	Formula
Term Frequency (TF)	$\sum_{i=1}^{K} P_i(w)$
TF-IDF	$TF(w) \times \log \frac{K}{DF(w)}$
Domain Consensus	$-\sum_{i=1}^{K} (P_i(w) \times \log P_i(w))$
Maximum TF	$\max_i P_i(w)$
Term Score (TS)	$\sum_{i=1}^{K} P_i(w) \log \frac{P_i(w)}{(\prod_{i=1}^{K} P_i(w))^{\frac{1}{K}}}$
TS-IDF	$TS(w) \times \log \frac{K}{TF(w)}$
Maximum TS	$\max_i TS_i(w)$

Table 2: Topic-based features

We also used 6 single-topic document features (documents are regarded as separate topics). In fact, we used all above-mentioned topic-based features except Domain Consensus, since this feature is already considered in the section of traditional frequency-based features (cf. section 4.1).

4.3 Other Features

Other features considered in our study include:

- 5 Linguistic features: *Ambiguity* (determines whether the term candidate has multiple initial forms or may belong to multiple parts of speech), *Novelty* (determines whether the term candidate is described in morphological dictionaries), *Specificity* (determines whether the term candidate exists in the reference collection), *Nouns* (determines whether the term candidate consists of only Nouns), and *Adjectives* (determines whether the term candidate contains Adjective).

- Features for term candidates that play subject syntactic role in sentences, features for term candidates beginning with a capital letter, and features for term candidates beginning with a capital letter that do not start sentences. We consider 6 features for each such group (and thus 18 features in the whole): namely, *Term Frequency, Document Frequency, TF-IDF, TF-RIDF, Domain Consensus*, and *corpus-based TF-IDF*;

- 2 features for term candidates that are in the context window of the several most frequent predefined ones: *NearTermsFreq,*

NearTermsFreq-IDF (Nokel et al., 2012). *NearTermsFreq* is defined as the number of the term candidate occurrences in the context window of the several predefined most frequent words.

- Average position of *the first occurrence in documents*, and *Term Length*.

Thus, 27 features belong to this group. To sum up, the full list of features comprises 69 features for single-word candidates and 88 features for two-word term candidates.

5 Experiments

We studied models for single-word and two-word term extraction from two above-described corpora: Russian banking electronic magazines, and English part of parallel corpus Europarl.

To extract single-word and two-word term candidates from these corpora, documents were processed by morphological analyzers. Thus, for English corpus we used Stanford POS tagger (`http://nlp.stanford.edu/software/corenlp.shtml`), while for Russian corpus we used our own morphological analyzer. Besides, from the set of extracted English term candidates we excluded words from the stop list created for the experiments (*other, another, that, this, those, mrs, sir, etc.*), and word pairs including stop-words were excluded as well.

Having extracted term candidates, we trained combined models comprising the above-described types of term features. The features were combined by Gradient Boosting machine learning algorithm, which proved to be the best one in our study. Namely, we used an open-source realization of this algorithm from `http://scikit-learn.org`. It is well-known that Gradient Boosting has a lot of parameters that need to be tuned. So, in all experiments we fixed all parameters, except the number of trees and maximum allowed depth of trees, that were tuned in each experiment individually. Besides, for training and evaluation four-fold cross validation was applied, which means that every time the training set was three-quarters of the whole list while the testing set was the remaining part.

A term extraction model has to find the best order, where real terms should be located at the beginning of the ordered list of term candidates. As an evaluation measure, we used Average Precision (AvP) often applied as a measure for term extraction (Zhang, 2008), (Bolshakova et al., 2013). It is defined for a set D of all term candidates with a subset of approved ones $D_q \subset D$ as follows:

$$AvP(D) = \frac{1}{|D_q|} \sum_{1 \le k \le |D|} (r_k \times (\frac{1}{k} \sum_{1 \le i \le k} r_i)) \quad (2)$$

where $r_i = 1$ if the i-th term $\in D_q$ and $r_i = 0$ otherwise.

At the first step of experiments we separately studied term extraction models for single-word and two-word terms. As baselines we considered several well-known features: *Weirdness, TF-IDF, C-Value* for single-word models and *TF-IDF, C-Value, Mutual Information* for two-word ones. In the Figures 1, 2, 3, 4 plots of AvP on various numbers of most frequent candidates are presented for

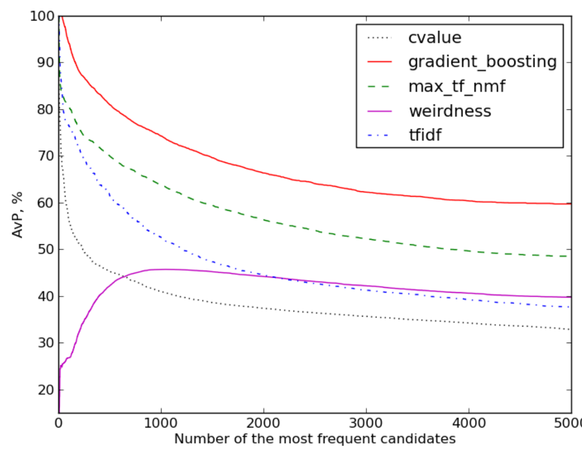

Figure 1: AvP for single-word Russian model

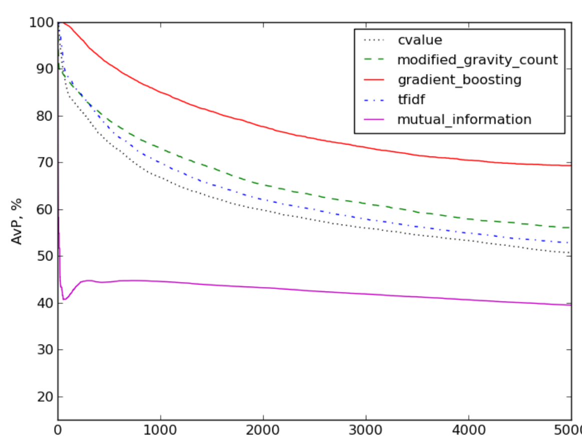

Figure 2: AvP for two-word Russian model

these baselines, the best single feature and the resulted model combined by Gradient Boosting.

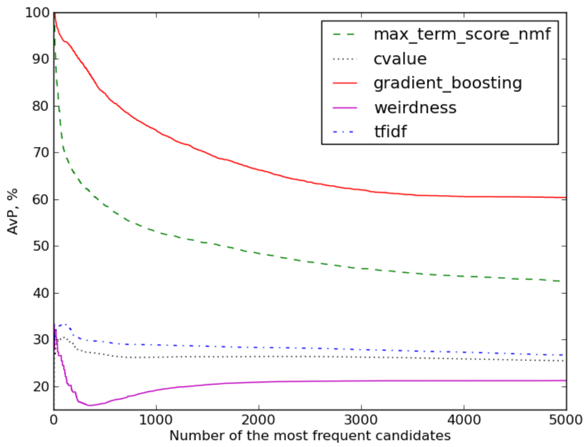

Figure 3: AvP for single-word English model

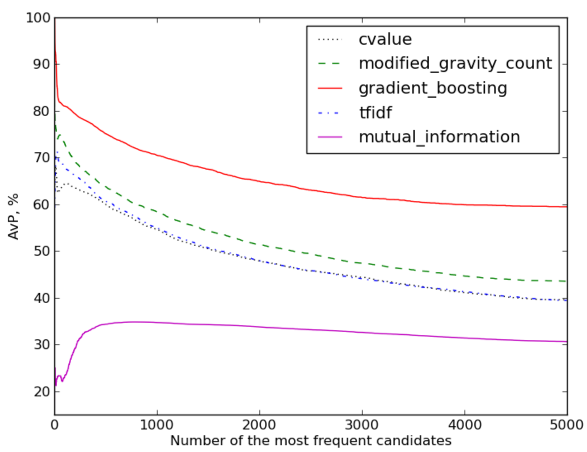

Figure 4: AvP for two-word English model

As we can see, the best single feature for single-word terms turned out to be a topic-based feature (either *Maximum Term Score* or *Maximum Term Frequency*), the performance of these features is considerably better than well-known baselines. So it seems that for single-word terms their relation to a domain subtopic is important.

The best single feature for two-word terms was found to be the novel context-based feature *Modified Gravity Count*. Besides, in all cases we can see the huge improvement of the combined model performance compared to well-known baselines and best single features.

At the second step of experiments we tried to determine the contribution of each above-described group of features to the whole combined model. We fixed the number of most frequent term candidates to 5000, excluded each of the following groups separately from the whole list: frequency-based features; features, based on the reference corpus; word association measures; context-based features, and topic-based features. The results of combining the remaining features by Gradient Boosting are presented in the Table 3.

Excluded group	Average Precision (%)			
	Single-word model		Two-word model	
	Russian	*English*	*Russian*	*English*
No (All features)	59.7	60.4	69.3	59.5
Frequency-based	59.5	59.6	68.9	58.6
Context-based	59.3	56.8	68.9	58.8
Reference corpus	57.5	59.6	68.3	55.6
Topic-based	56.8	59.4	68.9	60
Word association	–	–	69.3	59.7

Table 3: Contribution of feature groups to term extraction models

As we can see, features, which are based on the reference corpus, give the most significant contribution to the two-word term extraction models regardless of the subject domain and language.

Besides, the use of word association measures does not improve the quality of extraction of two-word terms. The latter conclusion contradicts the assumption of numerous studies that association measures should be useful for multi-word term extraction (Zhang, 2008), (Daille, 1995), (Kurz and Xu, 2002). From the other side, this conclusion can be quite evident because, for example, EuroVoc includes a lot of terms looking as compositional phrases with free separate usage of components (as *European party, European idea, economic consequence* etc.). Introduction of such terms into an information-retrieval thesaurus is possible due to multiple principles of term inclusion in information-retrieval thesauri (Z39.19, 2005).

At the last step of experiments we investigated both models for single-word and two-word term candidates together. We created a *unified model* for both types of term candidates, taking into account all features except association measures and obtaining as a result the *unified list* of candidates.

Then we created specific models separately for single-word and two-word term candidates. As the models are specialized, they can be potentially more efficient. We summed up resulted lists of extracted terms according to their probability values generated by Gradient Boosting, and in such a way obtained the *summed-up list* of term candidates. We should notice that in the case of the unified model there is more data to train it, so this model can be potentially very efficient too.

The comparison of AvP for these two models (for both corpora) shows that *summed-up model* slightly outperforms the *unified* one – cf. Figure 5.

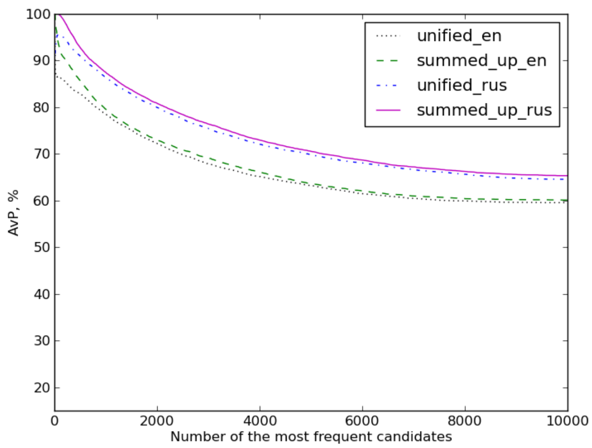

Figure 5: Unified vs summed-up models

In addition, as an example of the extracted term candidates, we present in the Table 4 the first 10 elements from the top of the term candidates lists created by unified models for Russian and English corpora (the elements in italics are real terms).

#	Russian corpus	English corpus
1	*Currency*	*Iran*
2	*Reporting period*	*Pakistan*
3	*Bond*	*Georgia*
4	Association	*India*
5	*Taxable period*	*Serbia*
6	*Reserve*	White paper
7	*Corporate governance*	*Syria*
8	*Credit history*	*Libya*
9	*Deal*	*Afghanistan*
10	*Borrower*	Member state

Table 4: Examples of term candidates extracted by unified models

The resulting unified models may be too complex in the number of applied features. Some of them may be redundant for Gradient Boosting and have no use in the models, make their training harder. In order to exclude them we applied a stepwise greedy *algorithm Add* for selecting the most significant features.

The algorithm starts with the empty set of features, and then at each step it adds the feature that maximizes the overall Average Precision, until there is any improvement between successive iterations. As a result, the combinations of only **13** features (out of total 69 features) were found for both corpora (see Table 5). We grouped similar features in the same rows of the table.

#	Russian corpus	English corpus
1	TF-RIDF Subjects	TF-RIDF Subjects
2	MGCount	MGCount
3	Lexical Cohesion	Lexical Cohesion
4	Nouns	Nouns
5	First Occurrence	First Occurrence
6	Weirdness	Weirdness
7	Corpus-based TF-IDF Non-Initial Words	TF-IDF Non-Initial Words
8	Sum3	Sum10
9	Term Score NMF	Maximum Term Score NMF
10	Single-topic TF-IDF	Single-topic Term Score
11	TF-RIDF	NearTermsFreq-IDF
12	KF-IDF	Term Variance Quality
13	TF-IDF NMF	Document Frequency

Table 5: Results of feature selection for unified models

Since there are representatives of all above-described groups in both found subsets of features, we conclude that each such group is significant for unified models of term extraction regardless of the scope and language. Besides, we can see that short models for both thesauri are quite similar.

6 Conclusion

In this paper we modelled single-word and two-word term extraction for the specific type of terminological resources – information-retrieval thesauri. Our experiments revealed features significant for extraction of single-word and two-word terms in the broad EuroVoc and relatively narrow banking domains. We showed that the best features for single term extraction in both cases are relatively new topic-based features, based on preliminary clustering of words in the target text collection. The context-based features are the most important for two-word term extraction.

The interesting result of our study is that the use of association measures does not improve the quality of term extraction models intended for information-retrieval thesaurus construction. It was also proved that the unified model can be applied to both single-word and two-word term extraction.

Acknowledgements

The work is partially supported by Dmitrii Zimin Dynastia Foundation with financial support of Yandex founders.

References

Ahmad K., Gillam L., Tostevin L. 1999. *University of Surrey Participation in TREC8: Indexing for Logical Document Extrapolation and Retrieval (WILDER)* Proceedings of TREC 1999.

Aze J., Roche M., Kodratoff Y., Sebag M. 2005. *Preference Learning in Terminology Extraction: A ROC-based Approach*. Proceedings of ASMDA'05, 209–219.

Basili R., Moschitti A., Pazienza M., Zanzotto F. 2001. *A Contrastive Approach to Term Extraction*. Proceedings of the 4th Terminology and Artificial Intelligence Conference.

Blei D. and Lafferty J. 2009. *Topic Models. Text Mining: Classification, Clustering and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

Bolshakova E., Loukachevitch N., Nokel M. 2013. *Topic Models Can Improve Domain Term Extraction*. Proceedings of ECIR 2013.

Bouma G. 2009. *Normalized Pointwise Mutual Information in Collocation Extraction*. Proceedings of the Biennial GSCL Conference, 31–40.

Church K. and Gale W. 1995. *Inverse Document Frequency IDF: A Measure of Deviation from Poisson*. Proceedings of the Third Workshop on Very Large Corpora. MIT Press, 121–130.

Church K. and Hanks P. 1990. *Word Association Norms, Mutual Information, and Lexicography*. Computational Linguistics, vol. 16, 22–29.

Daille B. 1995. *Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering* PhD dissertation. University of Paris, Paris.

Daudaravičius V. and Marcinkevičiené R. 2005. *Gravity Counts for the Boundaries of Collocations*. Corpus Linguistics, 9(2): 321–348.

Dunning T. 1993. *Accurate Metrics for the Statistics of Surprise and Coincidence*. Computational Linguistics, 19(1).

Foo J. and Merkel M. 2010. *Using Machine Learning to Perform Automatic Term Recognition*. Proceedings of the LREC 2010 Acquisition Workshop, Malta.

Frantzi K. and Ananiadou S. 1994. *The C-Value/NC-Value Domain-Independent Method for Multi-Word Term Extraction*. Journal of Natural Language Processing, vol. 6, no 3, 145–179.

Gelbukh A., Sidorov G., Lavin-Villa E., Chanona-Hermandez L. 2010. *Automatic Term Extraction using Log-likelihood based Comparison with General Reference Corpora*. Proceedings of the Natural Language Processing and Information Systems, 248–255.

Kurz D. and Xu F. 2002. *Text Mining for the Extraction of Domain Retrieval Terms and Term Collocations*. Proceedings of the International Workshop on Computational Approaches to Collocations.

Lee D. and Seung H. 2000. *Algorithms for Non-Negative Matrix Factorization* Proceedings of NIPS, 556–562.

Lopes G. and Silva J. 1999. *A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units*. Proceedings of the 6th Meeting on the Mathematics of Language, 369–381.

Loukachevitch N. 2012. *Automatic Term Recognition Needs Multiple Evidence*. Proceedings of LREC'12.

Liu L., Kang J., YU J., Wang Z. 2005 *A Comparative Study on Unsupervised Feature Selection Methods for Text Clustering*. Proceedings of NLP-KE'05, 597–601.

Nakagawa H. and Mori T. 2003. *Automatic Term Recognition Based on Statistics of Compound Nouns and their Components*. Terminology, vol. 9, no. 2, 201–219.

Nokel M., Bolshakova E., Loukachevitch N. 2012. *Combining Multiple Features for Single-Word Term Extraction*. Proceedings of Dialog 2012, 490–501.

Park Y., Byrd R., Boguraev B. 2002. *Automatic Glossary Extraction: Beyond Terminology Identification*. Proceedings of the 19th International Conference on Computational Linguistics.

Peñas A., Verdejo V., Gonzalo J. 2001. *Corpus-based Terminology Extraction Applied to Information Access*. Proceedings of the Corpus Linguistics 2001 Conference, 458–465.

Sclano F. and Velardi P. 2007. *TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities*. Proceedings of the 7th Conference on Terminology and Artificial Intelligence.

Smadja F., McKeown K. R. and Hatzivassiloglou V. 1996. *Translating Collocations for Bilingual Lexicons: A Statistical Approach*. Computational Linguistics, 22(1), 1–38.

Vivaldi J., Marquez L, Rodriguez H. 2001. *Improving Term Extraction by System Combination using Boosting*. Proceedings of the 12th European Conference on Machine Learning, 515–526.

Wong W., Liu W., Bennamoun W. 2007. *Determining Termhood for Learning Domain Ontologies using Domain Prevalence and Tendency*. Proceedings of the 6th Australasian Conference on Data Mining, 47–54.

Zhang Z., Iria J., Brewster C., Ciravegna F. 2008. *A Comparative Evaluation of Term Recognition Algorithms*. Proceedings of LREC 2008.

Zhang Z., Yoshida T., Ho T. B., Tang X. 2008. *Augmented Mutual Information for Multi-Word Term Extraction*. International Journal of Innovative Computing, Information and Control, 8(2), 543–554.

Z39.19 2005. *Guidelines for the Construction, Formal and Management of Monolingual Thesauri* NISO.