Linguistic Frontiers • 4(2) • 2021
DOI: 10.2478/lf-2021-0014

Linguistic **Frontiers**

Machine Learning in Terminology Extraction from Czech and English Texts

Original Study

Dominika Kováříková
Institute of the Czech National Corpus. Charles University, Faculty of Arts.

E-mail: dominika.kovarikova@ff.cuni.cz

Received: July 2020; Accepted: February 2021

Abstract: The method of automatic term recognition based on machine learning is focused primarily on the most important quantitative term attributes. It is able to successfully identify terms and non-terms (with success rate of more than 95 %) and find characteristic features of a term as a terminological unit. A single-word term can be characterized as a word with a low frequency that occurs considerably more often in specialized texts than in non-academic texts, occurs in a small number of disciplines, its distribution in the corpus is uneven as is the distance between its two instances. A multi-word term is a collocation consisting of words with low frequency and contains at least one single-word term. The method is based on quantitative features and it makes it possible to utilize the algorithms in multiple disciplines as well as to create cross-lingual applications (verified on Czech and English).

Keywords: term extraction, automatic term recognition, machine learning, corpus data, term characteristics

INTRODUCTION

Automatic term recognition (ATR) studies are usually focused on high success rate and the main goal is as accurate and successful an extraction of terms as possible. Authors use combinations of various qualitative and quantitative features of terms and non-terms in texts to achieve this primary goal.

This study[1] is based on automatic term recognition (ATR) but concentrates primarily on new knowledge about terms and their attributes, rather than solely on successful extraction. The main goal is not the success rate but finding out which of the term characteristics are the most relevant in the process of the extraction of terms. Such characteristics can help describe terms from a new point of view and might even become a part of a term definition. Evaluation of term features and the effectiveness of their combinations in term mining can be a resource for other ATR methods. Of course, for the results to be reliable, the method needs to be quite successful (high precision and recall), but it is not the research priority and main goal.

The machine learning[2] algorithms for term extraction as well as the evaluation of the features is provided by the data mining tool Weka (Waikato Environment for Knowledge Analysis, Hall et al. 2009). Data mining is defined as a (semi)automatic process of discovering patterns in substantial quantities of data (Witten, Frank 2005, 5). The requirement for a "useful" pattern is that it allows us to make nontrivial predictions on new data—for example, we can predict which words in a text are terms and non-terms based on a specific combination of the attributes.

[1] This study was written within the programme Progres Q08 Czech National Corpus implemented at the Faculty of Arts, Charles University. It was also supported by the European Regional Development Fund—Project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734).
[2] In some contexts, machine learning and data mining are partially overlapping terms. We recognize, that the prevailing terminological preference for data mining in this article is just a matter of point of view.

The research has two versions: Czech and English. The original extensive research (Kovářiková 2017) was focused on terms in Czech academic texts. As a result, a small set of the most important features was identified for successful extraction of single-word terms (the process is explained in "Effective Set of Term Characteristics"). The simpler English version of the research was then based on an assumption that the terms will, from a quantitative point of view, behave similarly in other languages as well (e.g. similar distribution in academic disciplines or throughout a corpus etc.).

AUTOMATIC TERM RECOGNITION

There have been dozens of studies of automatic term recognition (or automatic term extraction, computer-assisted term acquisition etc., Yang 1986; Kageura, Umino 1996; Heid 1998/1999; Gamper, Stock 1998/1999; Lossio-Ventura et al. 2014; Nazar 2016) since the late 1980s. The cooperation of linguists, terminologists and computer scientists brings more and more successful and accurate methods for identifying terms in written texts. The methods use various statistical and linguistic characteristics of terms and non-terms, such as frequency, distribution, POS classification and many other properties to find as many terms in texts as possible.

An effective method of ATR is a basis for many applications, such as human or machine translation, automatic indexing of texts, dictionary construction, text type characterization etc. (Yang 1986; Lauriston 1995; Kageura, Umino 1996; Chung 2003). The concerns of ATR are primarily practical and not theoretical (Lauriston 1995). However, Kageura and Umino (1996, 18) point out that the results of ATR methods can be useful for a theory of terminology—the characteristics specific to terms might become a part of term description or definition.

ATR methods use mainly quantitative (statistical) features such as frequency in different texts, types of texts or in various disciplines, distribution in disciplines or in the corpus as a whole, or collocational characteristics (Yang 1986; Kaguera, Umino 1996; Chung 2003; Wermter, Hahn 2005; Kit, Liu 2008). Other researchers use linguistic properties as well, e.g. morphological or syntactical behavior, POS classification, stop-lists, etc. (Frantzi, Ananiadou 1997; Ville-Ometz et al. 2007).

CHARACTERISTICS OF TERMS

Many researchers noticed that the terms have specific quantitative features within texts (Bečka 1972; Yang 1986; Kageura, Umino 1996; Chung 2003). Most of them can be divided into three main categories:

1. Frequency (e.g. relative frequency in a discipline or in non-academic texts, risk ratio—ratio of relative frequency in a discipline and in a reference corpus)
2. Distribution (e.g. distribution in various disciplines)
3. Contextual characteristics (e.g. entropy of the immediate left or right context)

Some of the researchers are convinced that quantitative characteristics alone are capable of distinguishing terms from non-terms. Bečka (1972) claims that "words with terminological validity may as lexical components be characterized in quantitative terms". According to Yang (1986), it is possible to identify terms "on basis of their frequencies of occurrence and distribution" (Yang 1986).

One of the main motives for that is the use of the data mining tool which requires computer processable input. However, it is not the only reason. The other advantage of using the quantitative features is relatively easy access to the information—unlike part-of-speech categorization or "internationality" of a word, it is possible to automatically calculate frequencies of words without extensive linguistic knowledge. Also, as was confirmed by the English version of the research, the quantitative features of terms are transferable among at least some languages to a certain extent (in opposition to the length or structure of a word).

Frequency:	
RFQdisc	relative frequency in a discipline
RFQsci	relative frequency in the subcorpus of academic texts (SCI)
RFQref	relative frequency in the reference subcorpus of non-academic texts (**REF**)
RFQdiscRFQref	risk ratio—relative frequency in a discipline to a relative frequency in REF
RFQsciRFQref	ratio of relative frequency in SCI to a relative frequency in REF
NoRef	the word does not occur in REF
Distribution:	
RDist	relative distribution in all available disciplines
SDRFQ	standard deviation of relative frequency of the word in all available disciplines
ARF	average reduced frequency—hows evenness of distribution throughout the corpus (Savický, Hlaváčová 2003) as well as frequency of the word in corpus
RARF	relative average reduced frequency—ARF divided by the frequency
SDRD	standard deviation of relative distance of two neighboring occurrences of the word
Context:	
ContextE	entropy of the immediate left and right context of the word in the corpus
ContextEr1	entropy of the immediate right context
ContextEl1	entropy of the immediate left context
Hgen	weighted average of relative frequencies of the preceding context (Hgen1 to Hgen5)
Linguistic features:	
Lensyl	length of the word (in syllables)
Struct	structure of the word – how usual or unusual the structure is (Greek and Latin words are different in structure from Czech ones)

Table 1: Candidate features of terms

METHOD AND MATERIAL

DATA MINING

Data mining is able to process and classify data of substantial quantities by computer algorithms. One way to use the data mining methods is to train them to find a specific combination of features that separate one group of words from another (in the case of linguistic data), for example, terms from non-terms. For that, data mining methods need a specific training input, where the terms and non-terms are manually labeled. After the learning process, the methods are able to find terms and non-terms in any data (providing the data contain all necessary information, primarily the values of individual features, such as frequency and distribution in texts).

Data mining has two substantial advantages: 1) it is able to track complex non-linear relations between individual term characteristics, and 2) it has the capacity to identify the most relevant of the examined features (feature ranking and feature selection).

The data mining tool used in this research is Weka (Waikato Environment for Knowledge Analysis, Hall et al. 2009). It assembles a group of algorithms for data analysis and predictive modeling. It supports standard data mining tasks: classification, regression, clustering, feature ranking and selection. Weka offers the opportunity to work with a number of methods which can be chosen for a specific research assignment—the individual methods can be compared based on their performance on the assignment. The method with the best performance in term extraction was Bagging-PART (based on the rule-based method PART, it combines decisions of several different models) which had the highest success rate (Šrajerová et al. 2009).

Data mining used for term extraction is a complex process. First, the data (words and their features) need to be prepared, then the available methods are trained to recognize the terms and non-terms based on the features. The resulting model is able to identify terms in any new text. During the process, it is possible to rank the term characteristics according to their importance for the term extraction or find a smaller set of very important features.

The presented research involved several distinct steps:

STEP 1: TRAINING DATA PREPARATION

The training data for this particular task contain several thousand words from Czech academic texts. Values of individual features (table 1) are automatically added to each of the words. For the purposes of the training process, every word is manually labeled as a term or a non-term.

STEP 2: TRAINING AND CROSS-VALIDATION

A method (or several methods[3]) from the Weka data mining tool is trained to find an algorithm which is able to distinguish terms from non-terms based on values of individual term characteristics. For example, there is a high probability a word is a term if it occurs only in one discipline, its relative frequency in the discipline compared to a non-academic corpus is high and it is non-evenly distributed in the whole corpus.

To evaluate the results of the training process and to avoid biased results, the data mining methods within Weka use cross-validation.

STEP 3: RANKING OF FEATURES

Some of the methods in the Weka data mining tool are able to evaluate features by their performance in the term extraction process. The ranking of the features is a very important result of the presented study: it tells us which of the characteristics of terms are the most important or typical and thus can be used in other ATR methods and in a description of terms in general (more about ranked features in "Ranking of Term Characteristics by Importance").

STEP 4: SELECTION OF AN EFFECTIVE SET OF FEATURES

The full set of features is not needed for the identification of terms and non-terms. The smaller the set of features necessary for term extraction, the more transparent is the description of a term based on it. Also, the preparation of testing data is much simpler and faster if there is only a small number of attributes calculated for each word (see "Effective Set of Term Characteristics").

STEP 5: ALGORITHM BASED ON THE SET OF CHOSEN FEATURES

An algorithm can be created on the basis of the smaller set of features identified in the previous step—the advantage is simpler testing data preparation. The other possibility is to use the original algorithm from step 2.

STEP 6: TESTING DATA PREPARATION

Testing data can consist of any text selected for analysis. The form of the data corresponds to the form of the training data—words with values of the features added automatically. The testing data can be diverse and extensive; therefore it is advantageous to calculate a smaller number of features rather than a larger one (see step 4).

STEP 7: AUTOMATIC TERM AND NON-TERM RECOGNITION IN TESTING DATA

Term candidates are automatically identified in the testing data. Each word is assigned a value from 0 to 1, where 0 is the strongest non-term and 1 is the strongest term. The boundary between term candidates and non-terms is set at 0.5 which is a default value in Weka and has also been experimentally verified as acceptable for this specific term mining task (Kovářiková 2017).

CORPORA

The corpus used for the research of the terms in Czech is a corpus of contemporary written Czech, SYN2010,

[3] The complete list of used methods as well as their success rate is listed in Kovářiková (2017, 65—68).

which is part of the Czech National Corpus. It contains 122 million words (with punctuation); 40 % of the corpus is fiction, 27 % is non-fiction (including technical, professional and academic texts) and 33 % is journalism. The English corpus en_syn that was created specifically for this research is similar to the SYN2010 in design but is not as large: it has 17 million words (including punctuation). It contains 40 % of fiction, 27 % of pseudo-academic texts from the English Wikipedia and 33 % of journalism. Similar proportions in the two corpora are very important for calculation of some of the term features (specifically ARF and SDRD).

Calculations of some other term characteristics depend on comparison of different types of texts, namely academic texts versus non-academic, non-professional texts such as fiction and journalism. For that reason, two subcorpora of each corpus were created: a subcorpus of academic texts named SCI and a subcorpus of fiction and journalism named REF. Table 2 shows the number of words in each subcorpus.

Name of Subcorpus	Lang.	Original Corpus	# of Words	# of disciplines
SCI1	CZ	SYN2010	9 million	37
SCI2	EN	en_syn	6 million	20
REF1	CZ	SYN2010	80 million	N/A
REF2	EN	en_syn	11 million	N/A

Table 2: Number of words in Czech and English subcorpora SCI (subcorpus of academic texts) and REF (reference corpus)

TRAINING AND TESTING DATA

The training data for the data mining tool contain several texts from four academic disciplines, as varied as possible (computers, literature, medicine, sociology). The total length of the texts is 8000 words, i.e. 2000 words for each discipline. Words are not lemmatized, all the instances in the research are word-forms.

Each word in the training data was manually labeled as a term or a non-term. For the purpose of the present study, a term is a word that can be found in a terminological dictionary of the given discipline. Problematic instances were decided by a specialist in the discipline.

A number of features presumed characteristic for terms in some way (table 1) was assigned to each of the words in the training data. The individual values of the features were calculated automatically.

The assumption is that terms have specific quantitative characteristics that are similar in all academic texts and that the algorithm trained on data from four different disciplines can be applied to texts from other fields or professions, and even texts in a different language to some degree.

All available academic texts from SYN2010 of the Czech National Corpus1 (9 million words from 37 disciplines) were prepared for automatic labeling as terms and non-terms: values of carefully selected features (see "Ranking of Term Characteristics by Importance") were assigned to each of the words. The same process was used for all (pseudo)academic texts from corpus en_syn (corpus in English, 6 million words from 20 disciplines). The testing data form manual evaluation of the automatic labeling consists of 37 small datasets (100 words) from each discipline in Czech and 20 datasets for English. Training and testing data are disjunctive.

POS TAGGING

One of the questions addressed in the research was usefulness of linguistic tagging, namely POS tagging, for the given assignment.

POS classification is assumed to be useful in term extraction since most of the single-word terms are nouns (Čermák 2010). It was included as one of the features used for term identification in an earlier stage of the study. However, the experiments suggested that POS tagging did not improve the results sufficiently enough to compensate for the drawbacks such as technical requirements for POS tagging in different languages (Kovářiková 2017).

EVALUATION OF THE RESULTS

The results of the data mining process are evaluated by standard statistical measures, i.e. precision, recall and accuracy (Manning, Schütze 2000). Term identification is a binary classification—all instances (words) are classified either as a term or a non-term. The evaluation is based on the number of terms that were correctly identified as terms (true positive, TP), the number of non-terms correctly classified as non-terms (true negative, TN), the number of terms incorrectly labeled as non-terms (false negative, FN) and the number of non-terms incorrectly identified as terms (false positive, FP).

Accuracy is a statistical measure that is able to assess the proportion of the correctly labeled words in the text (terms and non-terms). One hundred percent accuracy means that all the true terms were classified as terms and all the true non-terms were classified as non-terms. Precision is the ratio of the correctly identified terms to all words labeled as terms (correctly and incorrectly). Recall is the ratio of the correctly identified terms to all true terms (labeled as terms and as non-terms).

PROBLEMS OF TERMINOLOGICAL WORK

Many problems of automatic term recognition work are based on the fact that there is no clear boundary between terms and non-terms. Terminologists have observed that to manually distinguish terms from non-terms in a text is not an easy task (Bečka 1972). The value of terminological validity (Bečka 1972) of a word can be high or low—in simple terms, distinction between terms and non-terms is a matter of scale (Čermák 2010).

The position of a word on the scale is affected by various indicators, such as distribution of the word in academic disciplines, its frequent presence or total absence in a comparison corpus, termhood or technicalness of the term (how closely it is related to a particular discipline, Kageura, Umino 1996), frequency in the texts of the discipline, internationality of the word (Greek or Latin origin) or its length (especially in chemistry) or even its presence in a terminological dictionary. Some of the indicators are verifiable (terminological dictionary), and others can be detected intuitively to some degree or calculated accurately (frequency and distribution).

There is another barrier for terminological work, and it is the inability of one researcher to cover all knowledge necessary for full understanding of the examined subject—a complete set of all terms in all academic and professional disciplines. In many cases, there are two options: an extensive cooperation with experts in other fields of study, or a qualified estimate (or their combination).

In this respect, the first issue of terminological work is the classification of the academic and professional disciplines. For instance, the researchers need to decide whether to examine biology as one discipline or as zoology, botany and general biology. Another issue is categorizing disciplines as humanities, social sciences, natural sciences, applied or theoretical sciences, etc. Terminologists with a background in linguistics are often not trained experts on such questions, but still have to make decisions that affect their methods and results. Authors of the Editorial statement offer a practical solution to the problem: to acknowledge that the definition of a discipline (or a "subject field") is arbitrary and should be adjusted to the objectives of the individual research (L'Homme et al. 2003, 153). In the case of this project, the classification of academic disciplines is based on the classification proposed by the Czech National Corpus (structural attribute: genre).

SUCCESS RATES OF ATR METHOD BASED ON DATA MINING

The highest possible success rate of the method is not the only goal of this study—one of the main objectives is to identify the typical term characteristics. However, the better the results of the method are, the more reliable the conclusions based on it is. That is why we consider the high success of this method to be very important as well.

The success rates of the data mining technique were very high for the four disciplines chosen for training data in Czech (computers, literature, medicine and sociology): precision for data from all four disciplines was 89 %, recall almost 86 %, and accuracy (proportion of correctly labeled both terms and non-terms) 95 %. The terms were also automatically identified in all Czech academic texts available in the SYN2010 corpus. The texts consist of more than nine million words divided into 37 academic disciplines. The automatic extraction of terms was performed on English academic texts as well.

The success rate was measured on testing data—random short texts (one hundred consecutive words) from each of the subject fields. For Czech, the total of manually examined words was almost four thousand words, and two thousand words for English. For each word, the correctness of the automatic label (term/non-term) was evaluated.

The values of estimated accuracy, precision and recall in testing data for Czech academic texts are summarized in figure 1. Figure 2 shows the comparison of precision and recall in testing data in Czech and English academic texts (20 disciplines).

The figures show very high precision on average for Czech as well as English texts—average of about 85 %,

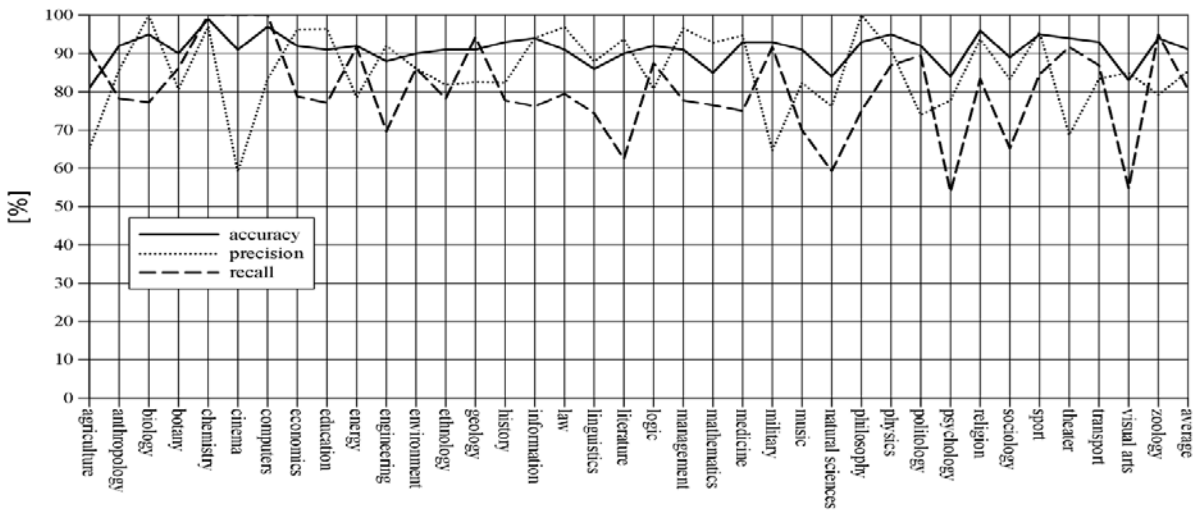

Figure 1: Accuracy, precision and recall of the ATR method in 37 academic disciplines (Czech texts only). The average values are on the right side of the graph.

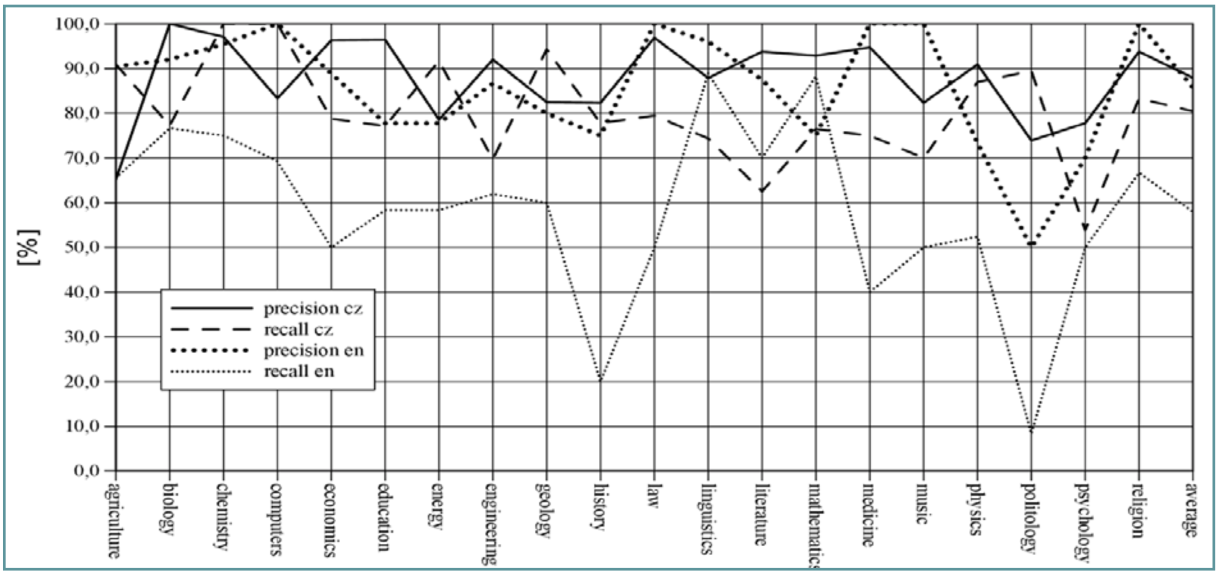

Figure 2: Precision and recall of the ATR method in 20 academic disciplines (comparison of Czech and English texts). The *average* values are on the right side of the graph.

with English disciplines being much more uneven. Recall in average in Czech texts is a little higher than 80 % and much lower in English texts: just below 60 %. Accuracy in Czech texts is in most cases 90 % or higher.

Since the testing data (100 words for each disciplines) were chosen randomly, it is possible that some of the samples are less suitable than others. Therefore, it is necessary to consider the results an estimate. Manual evaluation of the words labeled as terms and non-terms is based on available sources such as online terminological dictionaries, Wikipedia, etc.

RANKING OF TERM CHARACTERISTICS BY IMPORTANCE

Assessment of the distinctive term characteristics is provided by thirteen various methods available in the Weka data mining tools (for the complete list of the methods, see Witten, Frank 2005, 421). Primarily, such method components are intended to reduce the number of features and choose the most powerful ones so that the method can work more effectively with a smaller set of attributes. As a side effect, it offers the evaluation and ranking of the features with respect to their importance for the data mining process. It is safe to assume that such features are typical for terms and may be included in a description of a term.

All feature evaluation methods in Weka were used at the same time to produce the final feature rank. Each term attribute was assigned a normalized value between 0 and 1 by each of the methods depending on its significance or relevance rating by the method. The sum of the values determines the final ranking of the individual single-word term features that is displayed in figure 3 (the value on the y axis is a sum of the normalized value and does not have special significance).

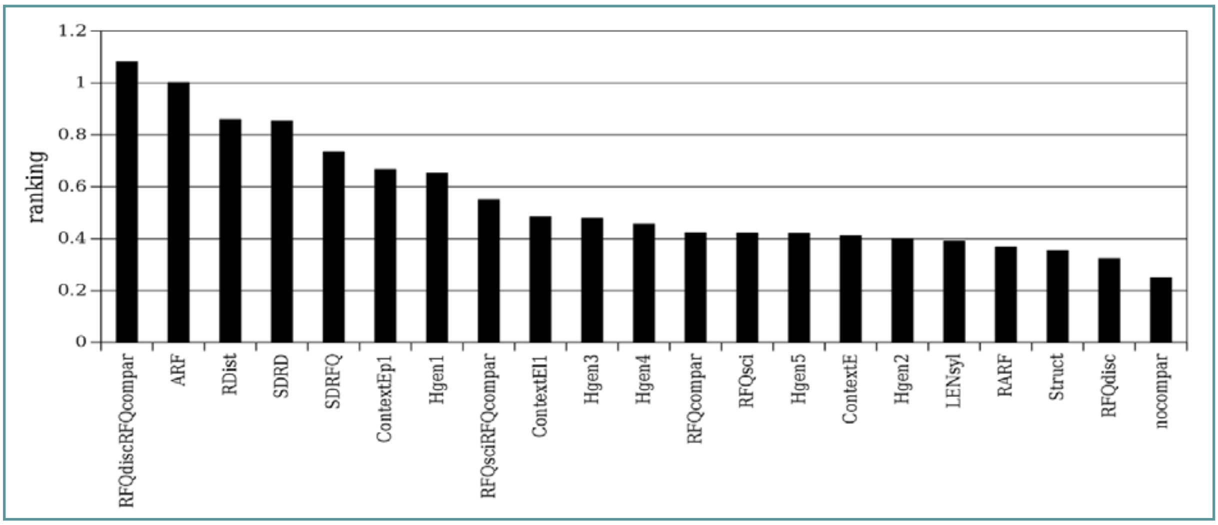

Figure 3: Feature ranking of all examined term features. (the value on the y axis is a sum of the normalized value and does not have special significance). The most significant features are on the left side.

The most important single-word term characteristic proved to be RFQdiscRFQref feature (risk ratio—relative frequency in a given discipline to relative frequency in the reference corpus). The second and third most influential characteristics are distributional features ARF (average reduced frequency) and RDist (relative distribution in disciplines). The feature ranked fourth, SDRD, is distributional as well (standard deviation of relative distance of two neighboring occurrences of the word).

The ranking itself does not contain information whether the attribute affects the identification of a word as a term in a positive or negative way. Such information can be provided by a correlation analysis of individual features with the fact that the word was manually or automatically labeled as a term. The results of the correlation analysis are shown in table 3. Positive correlation means that the higher is the value of the feature assigned to a word, the higher the probability the word is a term. Negative correlation (negative values) means that the lower is the value of the feature, the higher the probability the word is a term. Table 3 lists the features most strongly correlating with the word being labelled as a term.

Rank	Feature	Correlation with term. value
1	RFQdiscRFQref	0.6
2	RDist	-0.55
3	ARF	-0.49
4	SDRD	0.37

Table 3: Correlation of features and terminological validity of a word

RFQdiscRFQref
The higher is the risk ratio (relative frequency of a word in a discipline to a relative frequency in REF), the higher the probability the word is a term.

RDist
The lower the relative distribution in the disciplines, the higher is the probability the word is a term.

ARF
The less evenly is a word distributed throughout the corpus (lower average reduced frequency), the higher is the probability the word is a term.

SDRD
The less evenly is a word distributed throughout the corpus (higher standard deviation of relative distance of two neighboring occurrences of the word), the higher is the probability the word is a term.

EFFECTIVE SET OF TERM CHARACTERISTICS

For this study, a total of 17 features were examined in order to find the most important ones. The set of features is quite large—to find a smaller set of attributes effective enough for identifying terms would be more appropriate. Firstly, it would simplify the process of preparing the material for the ATR method. Also, it would show if a smaller number of the most important features was sufficient for term identification (and make it possible to discard the rest of the attributes).

To find a compact set of term characteristics, a number of combinations were tested by the Bagging-PART method available in Weka. The sequence of the features was given by the feature ranking (combinations of features 1+2, 1+2+3, 1+2+3+4 etc. were examined). The improvement of the results is shown in figure 4.

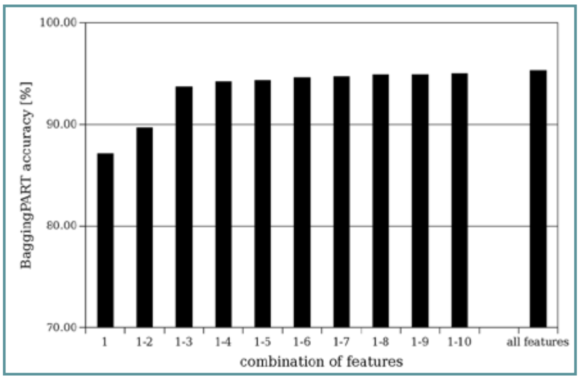

Figure 4: Comparison of a term identification success rate (accuracy measure) for combinations of term features

Based on figure 4, we can conclude that the best set of features for single-word terms includes four term characteristics: RFQdistRFQref (risk ratio), ARF (average reduced frequency), RDist (relative distribution in disciplines) and SDRD (standard deviation of relative distance in texts).

A single-word term then can be described as a word that can be found in academic texts substantially more often than in non-academic texts; it occurs only in a small number of disciplines; it is not very frequent and is spread unevenly throughout a corpus (such as SYN2010 or en_syn); and the distances between its individual occurrences are uneven.

THE TERMIT APPLICATION

The resulting lists of words automatically labeled as non-terms and terms (on a scale from 0 to 1) was utilized in an online application for term and discipline identification—TERMIT (www.korpus.cz/termit). The application identifies terms in any Czech text based on a list of terms automatically extracted from the academic texts available in corpus SYN2010.

According to the identified terms, TERMIT also indicates the prevailing academic discipline as well as other dominant disciplines. Most of the examined texts included terms from a number of academic disciplines—it does not mean that most of the texts are interdisciplinary, rather it suggests that academic disciplines share a very high number of frequently used terminology (such as *valency* in chemistry and linguistics, *case* in linguistics and law, or *communication* in linguistics and biology).

For English, beta version of the TERMIT application is also available.

CONCLUSION

The ATR method based on machine learning (data mining) has several strengths. First, it provides a high success rate in identifying terms in academic texts based on very complex relations between individual term characteristics. Also, it is able to assess the role of the features and thus provide the ranking of term characteristics. Based on the rank of the features, we are able to find a compact (small but effective) set of term characteristics which is important for two reasons: 1) such features can be used to describe a term from a quantitative point of view which was the main goal of this study, and 2) future ATR methods can find inspiration in such knowledge.

The fact that the terms can be identified based on quantitative features makes it possible to utilize the algorithms in multiple academic disciplines as well as to create cross-lingual or even multi-lingual applications (so far verified only on two languages: Czech and English).

The very quality that facilitates all the advantages of data mining in ATR is at the same time the main drawback of the method which is the complexity of the data mining process as well as the resulting algorithms.

Overall, the presented automatic term identification method based on data mining worked quite well considering its broad range (dozens of academic disciplines in two languages). An important future task is testing this ATR method on larger corpus data as well as determining whether the method is stable by means of statistical tests such as t-test, bootstrapping, or resampling.

The results of the study were utilized in an online application TERMIT for automatic term identification which is available in Czech as well as English version.

BIBLIOGRAPHY

Bečka, J. V., 1972. The lexical composition of specialized texts and its quantitative aspect. *Prague Studies in Mathematical Linguistics,* 4, 47—64.

Čermák, F. (2010). *Lexikon a sémantika*. Praha: NLN.

Křen, M. et al., 2010. *SYN2010: žánrově vyvážený korpus psané češtiny*. Institute of the Czech National Corpus, Charles University, Prague, available at: < http://www.korpus.cz >.

Chung, T. M., 2003. A corpus comparison approach for terminology extraction. *Terminology,* 9(2), 221—246.

Cvrček, V., 2013, *Kvantitativní analýza kontextu*. Praha: NLN/ÚČNK.

Frantzi, K. T., Ananiadou, S., 1999. The C/NC value domain independent method for multi-word term extraction. *Journal of Natural Language Processing,* 3(2), 115—127.

Gamper, H., Stock, O., 1998/1999. Corpus-based terminology. *Terminology,* 5(2), 147—159.

Hall, M. et al., 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11(1), 10—18.

Heid, U., 1998/1999. A linguistic bootstrapping approach to the extraction of term candidates from German text. *Terminology,* 5(2), 161—181.

Kageura, K., Umino, B., 1996. Methods of automatic term recognition: A review. *Terminology,* 3(2), 259—289.

Kit, C., Liu, X., 2008. Measuring mono-word termhood by rank difference via corpus comparison. *Terminology,* 14(2), 204—229.

Kováříková, D., 2017. *Kvantitativní charakteristiky termínů*. Praha: NLN/ÚČNK.

L'Homme, M., Heid, U., Sager, J. C., 2003. Terminology during the past decade (1994-2004): An editorial statement. *Terminology,* 9(2),151—161.

Lauriston, A., 1995. Criteria for measuring term recognition. In: *EACL '95 Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*. San Francisco: Morgan Kaufmann Publishers.

Lossio-Ventura, J. A. et al., 2014. Biomedical Terminology Extraction: A new combination of Statistical and Web Mining Approaches. *JADT'2014: Journées internationales d'Analyse statistique des Données Textuelle,* 421—432.

Manning, C. D., Schütze, H., 2000. *Foundations of Statistical Natural Language Processing*. Cambridge/ London: The MIT Press.

Nazar, R., 2016. Distributional analysis applied to terminology extraction. *Terminology,* 22(2), 141—170.

Savický, P., Hlaváčová, J., 2003. Measures of word commonness. *Journal of Quantitative Linguistics,* 9(3), 215—231.

Šrajerová, D., Kovářík, O., Cvrček, V., 2009. Automatic term recognition based on data-mining techniques. *Proceedings of Computer Science and Information Engineering—CSIE*. Los Angeles.

Ville-Ometz, F., Royauté, J., Zasadzinski, A., 2007. Enhancing in automatic recognition and extraction of term variants with linguistic features. *Terminology,* 13(1), 35—59.

Wermter, J., Hahn, U., 2005. Finding new terminology in very large corpora. In: *Proceedings of the 3rd International Conference on Knowledge Capture (KCAP 2005)*, 137—144.

Return to front page ↑