Automatic Term Recognition Based on Data-mining Techniques

Dominika Šrajerová
Institute of the Czech National Corpus
Faculty of Arts, Charles University
Prague, Czech Republic

Oleg Kovářík
Department of Computer Science and Engineering
Faculty of Electrical Engineering
Czech Technical University in Prague

Václav Cvrček
Institute of the Czech National Corpus
Faculty of Arts, Charles University
Prague, Czech Republic

Abstract

We present a new method for automatic term extraction which is based on training datasets created to build inductive models for term identification. Existing approaches employ simple statistical and linguistic rules designed merely ad-hoc and are unable to utilize complex relations of linguistic units. In contrast to those approaches, our method does not require such manually ascribed rules of extraction. The data for our research is taken from the Czech National Corpus which is lemmatised and morphologically tagged. Statistical information (frequency, distribution etc.) is generated automatically and thus the only expert contribution needed is to label terms in the training dataset. The data mining software creates models that perform the extraction without any further human input. Additionally, feature ranking can serve as valuable aid for understanding of the extraction process and its future development and in terminology research.

1. Introduction

Automatic term recognition (ATR) is a process of selecting elements in a corpus that are considered terms of the discipline which is the object of inquiry. The results are applicable in machine translation, automatic indexing and other types of automatic language processing as well as for the construction of terminological dictionaries [4]. Additionally, the ATR can serve as a resource for stylometry and theory of terminology, namely for defining of a 'term' as a central notion of terminology.

Current ATR techniques are focused on extraction terms on the basis of different features of the term, statistical (based on frequency, distribution in fields of study, etc.) and linguistic (parts of speech, morphological categories), that are used as criteria for term recognition in a text. The researchers usually select features that will be used for the given extraction method prior to their experiments (see [5, 7, 8]).

In contrast to such ATR techniques, our method is not directed at the terms themselves but rather at the criteria for selection of the terms. The data-mining program is provided with a substantial number of features which can possibly contribute to the specific 'essence' of a term. The significance of individual features is automatically detected within the procedure.

A term as a terminological unit is mostly defined as 'a conventional symbol that represents a concept defined within particular field of knowledge' [3] or 'the designation of a defined concept in a special language by a linguistic expression' [1].

It is obvious that such definitions are not sufficient for automatic extraction. One reason is that it gives us no formal or statistical description of the term that could be used for ATR. Furthermore, the definition of the concept represented by the term is often required which is in fact in contradiction to the automatic term recognition. Not all terms are defined, and those are the units ATR is aimed at (for example new terms in the discipline).

Although the term is not defined sufficiently, specialists in the given discipline seem to intuitively recognize words that are associated with their field of study. Based on this knowledge, we have prepared a new method for automatic term extraction which is based on training datasets created to build inductive models for term identification. The only expert contribution needed is the labeling of terms in the training dataset. The data mining software FAKE GAME creates models that perform the extraction without any further human input, on the basis of automatically generated

statistical and linguistic information. Additional information about the extraction process useful for the further development of the ATR method and for the theory of terminology is provided by the built-in feature ranking.

2. FAKE GAME

FAKE GAME [6] is an open source data-mining tool developed at CTU, Prague. It is an extension of GMDH, a set of several algorithms for different solutions to problems. FAKE GAME constructs a special neural network on training data. The network is constructed of heterogeneous units with various transfer functions (linear, sigmoid, polynomial, etc.) layer by layer. In each layer, a large number of units that differ in transfer function and in the number of connections to the previous layer is generated. The best configuration of units (their transfer functions, internal parameters and connections) is then evolved by Niching Genetic Algorithm. In the next step, the units are evaluated on testing data and the least efficient units are deleted from the layer. The completed layer is 'frozen' and the algorithm continues by creating the next layer.

New layers are added until a unit with satisfactory output accuracy is found. After the training process is finished, the system extracts the equation from the model and this equation can be used on any data for which the appropriate variables were computed. Because the system enables us to choose which type of units we want to use, we can regulate the complexity of the resulting equation, e.g. by selecting only linear units. Outputs are from the interval $< 0, 1 >$ and for the classification purposes the threshold between classes is 0.5.

Furthermore, FAKE GAME can train several models and combine their responses in the process called ensembling to eliminate errors in individual models. Thus we can take equations from all models and calculate their average response to make use of the ensembling in our application. The response of a model is always a real number because all units expect real numbers as their inputs and they produce real value on its output. Due to this fact, all inputs are also taken as real numbers and it is necessary to convert nominal input attributes to 1 from N encoding which is the case with PO attributes (see section 3).

The accuracy of models can be measured in two ways. Data can be simply split to training and testing set and the accuracy of both of them can be measured; or automatic k-fold cross validation can be executed. The latter makes the results more reliable because it inhibits problems with training data sampling which can lead to incorrect accuracy estimation.

Very important for ATR is the ability to create the feature-ranking of the input attributes. FAKE GAME examines the structure of the model and estimates the importance of individual attributes. The reliability of this process can be again increased by creating an ensemble of models and averaging the results of feature ranking. This process is helpful in gaining knowledge about the data and reducing the number of input attributes which can lead to improvements in results.

3. Data

The data for our research are taken from the Czech National Corpus [2]. The corpus SYN2005 contains several types of texts. For the purpose of our study, the data were restricted to texts of academic literature, specifically texts of ten academic disciplines[1]. The number of compared disciplines is quite high in comparison to other ATR research projects, which is faciliated by the automatic processing of the data.

From the texts, training datasets of word-forms[2] were selected using number selection. Training dataset for each of the selected academic disciplines contains 1000 word-forms. The word-forms were manually marked as terms or non-terms, and were automatically assigned several statistical and linguistic features. Calculation of some of the features required a comparative corpus that contains non-scientific texts[3].

The list of features assigned to each word from each field of study is below.

FQ(disc) frequency of the word-form in texts of a given discipline

RFQ(disc) relative frequency of a word-form (i.e. frequency of the word divided by the total length of texts in a given field of study)

FQ(disc)/FQ(gen) frequency of a word-form in texts of a given discipline divided by frequency of the word-form in general non-scientific corpus

RFQ(disc)/RFQ(gen) relative frequency of a word-form in given discipline divided by the relative frequency of the same word in general corpus

Disc(only) boolean: 1 = word-form is restricted only to one field of study, 0 = word-form occurs in more than one discipline

Distr number of disciplines in which word-form occurs

[1] The chosen disciplines are: ART art history, BUI building industry, ECO economics, ENE power engineering, JUR jurisprudence, LIT literature, MED medical science, PHI philosophy, REL theology, ZOO zoology (sciences and social sciences, proper and applied).

[2] The decision to use word-forms rather then lemmas is substantial in Czech as it is an inflective language with rich morphology, and is based on the hypothesis that the meaning is connected more to a specific word-form than to lemmas. [9]

[3] List of text types used in the non-scientific comparative corpus: NOV novel, COL collection of short stories, VER poetry, SON songs, SCR scripts/film scripts, PUB journalism, ADM administrative texts, IMA other imaginative texts, MIS miscellanea (texts such as encyclopedias and textbooks were included neither in the data nor in the comparative corpus).

ARF(gen) reduced frequency; number of equal chunks of text in general corpus in which word-form occurs (number of the chunks is equal to the frequency of the word).

RR(disc) relative rank (i.e. rank of a word divided by the number of words in a given field of study)

Cover% coverage of texts in given field of study by sentences cointaing a given word-form

H(gen) average entropy of a word-form, calculated (using frequencies from general corpus) from a sequence of 5 preceding words

Len(syl) length of the word in syllables

Struct 'rareness' of structure of a word; sum of probabilities of each bigram in the word-form (probabilities were taken from words occuring in general corpus)

Case case of the first letter; U = lemma of the word begins always with upper-case letter (proper nouns), L = lemma of the word begins always with lower-case letter (common nouns), B = lemma of the word begins both with lower-case and upper-case letter.

PO list of parts of speech (more than one in case of homonymy). **PO** was substituted by 10 variables representing parts of speech occuring in Czech, i.e. **N** = nouns, **A** = adjectives, **P** = pronouns, **C** = numerals, **V** = verbs, **D** = adverbs, **R** = prepositions, **J** = conjunctions, **T** = particles, **I** = interjections and **I** = not recognized by morphology).

AvRPos-sen average relative position of a word-form in sentence in a given discipline

AvAbsPos-sen average absolute position of a word-form in sentence in a given discipline

Av1Pos-sen average position in sentence by the 1st occurance of the word-form in a document

Av1Pos-opus average position of the 1st occurance of the word-form in document

4. Method

The presented method aims to automatically recognize terms in academic texts. The advantage is that the procedure is almost fully automatic and human input is reduced to the marking of terms in the training dataset by a terminologist. A software developed to prepare the input for the data-mining tool assigns statistical and linguistic information to each word in the training dataset based on texts available in corpus SYN2005. The data-mining tool FAKE GAME constructs model(s) on the classified data. The response of the model(s) can be expressed as an equation where individual variables correspond to the statistical and linguistic information added in the previous step. The complexity of the equation is adjustable because the researcher can decide which type of units will be used for the calculation (e.g. linear units only) by the FAKE GAME. This equation can be used for automatic term recognition in any data with the according statistical and linguistic information.

Additionally, the feature-ranking points out the features that have the strongest impact on the automatic recognition of terms which can serve as a base for further refining of the original ATR method.

5. Experiments and results

For the purpose of accuracy evaluation we ran a set of experiments on the training data. In each experiment, we used a 10-fold cross validation (which is a built-in function of FAKE GAME) on a group of ten models. The resulting accuracy is expressed as a percentage of correctly classified instances.

The experiments are run on data with specific features. First, each word-form occurs only once in the data. It means that the most frequent words such as conjunctions or the verb 'be' have a smaller influence on accuracy then they would have in experiments with real text. And second, the words with frequency lower than 3 are removed from the data because there is a probability that they are typing errors.

The first round of experiments was targeted on the data from individual academic disciplines separately. The purpose was to detect any differences between the disciplines or between their groups (e.g. natural vs social sciences). Even though the individual disciplines vary in number of manually marked terms - the ratio of terms in the disciplines ranges from 24 % to 56 % of the monitored word-forms (see figure 1), no significant differences in the resulting accuracy of the ATR method have been observed.

In the second round, the data from individual disciplines were merged and the experiments were run with standard settings (heterogeneous units) as well as linear units only. Accuracy was measured with a full set of features. Again, no significant difference between the resulting accuracy for the separate and the merged data was observed.

For the third round of experiments the feature-ranking was used and so the significance of individual features for the computation was detected. The experiments were aimed at reducing the number of variables used as input information. Experiments run with a set of 8 features (out of the original number of 30 features) showed again no significant differences in the resulting accuracy. However, a test run with a set of 2 variables of the biggest influence (RFQ(disc)/RFQ(gen) and Distr[4]), none of which is linguistic information, demonstrated a lower accuracy than any of the experiments with the merged data from all the disciplines.

In figure 1 and figure 2 we show the final results of all experiments. Accuracy ranges from 80 % to 86 % in most

[4]For abbreviations see section 3.

Figure 1. Percentage of terms in datasets and average resulting accuracy in experiments

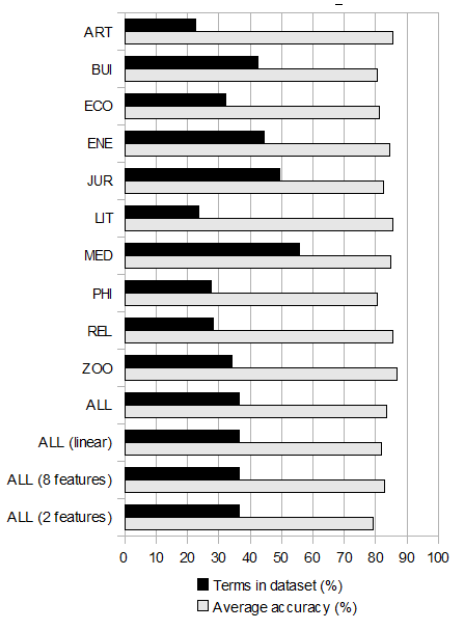
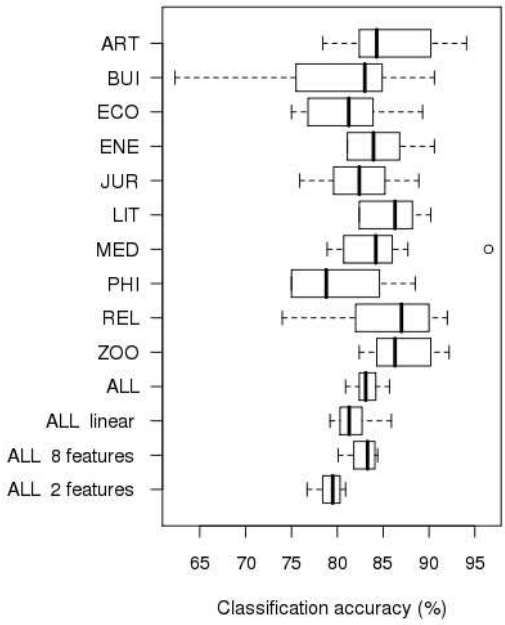


Figure 2. Resulting accuracy of ATR method in individual experiments



cases, with average of 83,2 %. The most obvious exception is a test with only 2 features. For the purpose of comparison, we calculated the accuracy of term extraction based on two most simple rules to determine the worst case performance. If all word-forms in the training data were classified as non-term, the accuracy of extraction would be 64 %, if all nouns were classified as terms, the accuracy would be 69 %.

It is necessary to remember that the data used for the experiments are specially adjusted - the results are supposed to be different for real text data.

We present two examples of the resulting equations: the linear (A) and the equation based standard settings of FAKE GAME (B). A word-form is considered a term if its score is higher than 0.5. The higher the score, especially with use of the linear equation, the more characteristics of a term are demonstrated by the selected word-form. In a real application, the threshold of 'termhood' can be raised to obtain more strict terms.

(A) $TERM = 0.21 \cdot N - 1.566\mathrm{E}^{-7} \cdot ARF + 0.236 \cdot RFQ(disc)/RFQ(gen) + 0.489 \cdot \textit{Case-L} - 0.123 \cdot V - 0.843$

(B) $TERM = 1/[1 + exp(-(6.750/(1 + exp(-(5.115/(1 + exp(-(13.232 \cdot (RFQ(disc)/RFQ(gen) - 0.01)/6.17 - 6.525))) + 3.450 \cdot \textit{Case-L} - 5.904))) + 1.724 \cdot N - 1.921 \cdot V - 3.962))]$

The equation (A) was for comparison reasons used for term extraction from a non-academic text. The text was a Czech translation of a the well known novel by Joseph Heller, *Catch-22*. Among the highest-ranked words selected on basis of the equation were mostly military and medical terms. This test proved that the ATR method works not only for academic texts that are considered the domain of terminology, but also for other texts such as a literary fiction.

6. Conclusions

We have presented a new method for automatic term recognition. In comparison to other ATR methods, the only expert (terminologist) input needed is for marking a sufficient number of terms in a text. All other steps in the method can be fully automated and these include the calculation of statistical and linguistic features for each word-form and the application of special data-mining software FAKE GAME. The outputs of the method are (1) extracted terms, (2) an equation that can be used on any text (with automatically added features) to extract terms and (3) an evaluation of the significance of the individual features.

The presented method assigns to each word a real number representing a degree of membership to the class of terms which can be used to adjust the strictness of term selection.

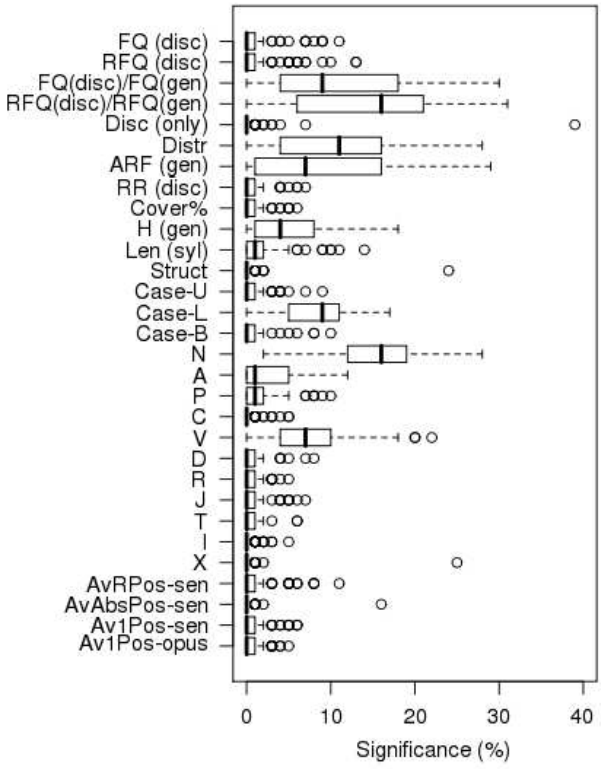
The experiments proved that the proposed ATR method was able to correctly recognize between 80 – 86 % of terms on the basis of marked terms and non-terms in training datasets. The datasets were created from academic texts by selecting one representative for each word-form and by

adding automatically generated list of features.

The method is also promising in the case of non-academic data such as fiction or other types of literature. Such data were until now of no interest for automatic term recognition researchers. However, the results of the application of our method to a non-academic text suggest that this might be attractive for terminologists in the future and might lead to a wider scope for term and terminology.

Within the described experiments, we did not observe any significant differences in the resulting accuracy of the method in the selected academic disciplines. It is surprising considering the differences between the individual academic disciplines (e.g. the difference in the term/non-term ratio). However, this result needs to be verified by experiments on larger datasets.

Figure 3. Significance of statistical and linguistic features for the ATR method



The results of the feature-ranking procedure indicate that the most influential features for our ATR method are (in order of significance): RFQ(disc)/RFQ(gen), PO: N, Distr, ARF, PO: V, Case-L, H, FQ(disc)/FQ(gen)[5] (see figure 3). This result corresponds with the intuitive linguistic perception of term.

[5]For abbreviations see section 3.

7. Future work

By carrying out more experiments on larger training datasets, we might be able to discover potential differences between disciplines or more precise significance of individual statistical and linguistic features of terms that could not be revealed due to limited amount of data.

Evaluation of the accuracy of ATR method in real texts is one of the most important next steps.

A step of similar importance is to search for a way of using the current method for finding multi-word terms as whole units.

The second area for further research is theory of terminology and linguistic theory. As mentioned above, our method proved able to extract terms from non-academic texts such as literary fiction. Terms included in texts that are traditionally not examined by terminologists might modify the conception of terminology.

More generally, the successful application of a data-mining tool on linguistic data suggests that a further exploitation of the data-mining for linguistic theory might be possible. Our method may be viable for automatically extracting linguistic phenomena other than terms on the basis of manual marking of training datasets.

References

[1] *Terminology work – Vocabulary – Part 1: Theory and application*. International Organization for Standardization (ISO).

[2] *Czech National Corpus – SYN2005*. Institute of the Czech National Corpus; http://www.korpus.cz, Prague, 2005.

[3] M. T. Cabré. *Terminology: Theory, Methods and Applications*. John Benjamins Publishing, Amsterdam, 1999.

[4] U. Heid and J. McNaught. Eurotra-7 study: Feasibility and. project definition study on the reusability of lexical and terminological resources in computerised applications. *Eurotra-7 Final Report*, 1991.

[5] K. Kageura and B. Umino. Methods of automatic term recognition: A review. *Terminology*, 3(2):259–289, 1996.

[6] P. Kordík. *Fully Automated Knowledge Extraction using Group of Adaptive Models Evolution*. PhD Thesis, Czech Technical University in Prague, FEE, Prague, 2006.

[7] I. Korkontzelos, I. Klapaftis, and S. Manandhar. Reviewing and evaluating automatic term recognition techniques. *Proceedings of the 6th International Conference on Natural Language Processing, GoTAL 2008*, pages 248–259, 2008.

[8] A. Lauriston. Criteria for measuring term recognition. *Seventh Conference of the European Chapter of the Association for Computational Linguistics*, pages 27–31, 1995.

[9] J. Sinclair. *Trust the text: Language, Corpus and Discourse*. Routledge, 2004.

[10] F. Čermák. Termín a frazém. *Termina 2000*, pages 31–36, 2000.