An approach to natural text classification using different types of classification features

A.V. Glazkova

Tyumen State University, 625003, 6 Volodarskogo street, Tyumen, Russia

Abstract

The article describes our approach to the task of text addressee detection. We consider four different types of classification features and suggest a step-by-step approach to text classification. In conclusion, the paper presents the results of the experiment that have been performed to check the classification accuracy using the proposed approach.

Keywords: informational retrieval; natural language processing; document; text classification; classification features; machine learning

1. Introduction

The problem of text classification is widely studied in the data mining, machine learning, database, and information retrieval communities with application in a number of diverse domains (such as document organization, content creation, spam filtering, web search, etc.).

Development of methods of ordering and retrieval of information is one of the key areas of modern computer science. The constant increase of information resources requires improved natural language text classification tools.

One of the pressing issues of text classification is a solution to the problem of determining the characteristics of the addressee. The problem was raised by D. Traum [1]. Since this paper only a few publications were devoted to the analysis of features that characterize the text in terms of its orientation to different categories of readers (N. Jovanovic [2], R. op den Akker [3], H. op den Akker [4]). Mentioned works were related to texts written in English and classification features were also selected for the English texts. Using these features for texts written in Slavic language is not correct due to the individual characteristics of the syntactic structure of each language. Currently, in most languages there is no single set of features which could be the basis for classification.

2. The object of the study

In this study, we consider the problem of classification by the example of text assignment to a particular age group of recipients. First of all, the ability to classify texts on the basis of age groups of their addressee improves the relevance of the results of information retrieval. Also it improves the mechanisms of exclusion undesirable resources from found sample (e.g., web-pages designed for the user from another category). The urgency of identifying the age of the recipient is justified by the presence of age restrictions on internet content resources in many countries, as well as the development of e-learning systems.

2.1. Classification feature types

Characteristics of natural language texts usually can be represented in different scales of measurement [5].

We consider the following types of classification features:

1) Binary ({0,1} e.g., presence / absence of special vocabulary in text).
2) Nominal (finite set of values e.g., literary form – story, novel; genre).
3) Ordinal (finite ordered set of values e.g., period of creation, audience education level).
4) Interval (interval value e.g., number of complex syntactic constructions; the number of sentences).

Some of binary, nominal and ordinal features cannot influence the text belonging to the category (for example, structure type text – prose or poetry). At the same time influencing features of these data types (binary, nominal, ordinal) may have different functions:

1) represent a marker limiting the number of categories for a text;
2) be a specific marker for the presence of additional clarifying features.

In particular, the presence of profanity in the text clearly indicates that the text does not address to the youngest readers. On the other hand, the binary feature characterizing the presence of illustration in the document requires image type clarifications. The text containing charts also probably cannot be related to the youngest audience.

2.2. Algorithm of text classification

The existence of markers allows step-by-step classification (fig. 1):

- Step 1. Evaluation of the influence of binary, nominal and ordinal features.
- Step 2. Estimation of the presence of feature values that clearly indicate the category of recipients or limit the number of categories for the text.

- Step 3. If the category of the text did not obtain at the second step, the category is determined on the basis of characteristic values, measured on an interval scale (the average length of sentences, the number of polysyllabic words, etc.).

We believe that this step-by-step approach will make the classification of texts represented by the multi-type features clearer and easier.

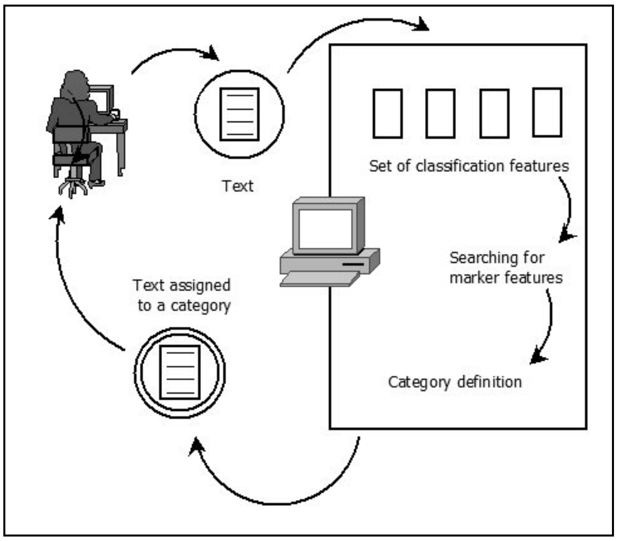

Text
Set of classification features
Searching for marker features
Category definition
Text assigned to a category

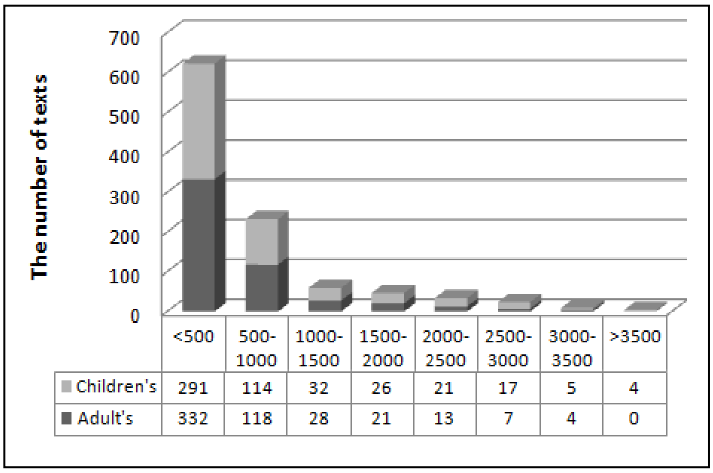
Fig. 1. The classification process scheme.

2.3. Corpus

In the computational experiments are used the Database “Morphological Standard of the Russian National Corpus” and “Database of meta tagging of the Russian National Corpus" (a collection of children's literature)” [6]. The Russian National Corpus [7] includes primarily original prose representing standard Russian. Each text in the corpus is subject to meta-tagging and morphological tagging.

The sample size is 532 texts of modern fiction (from the middle of the 20 century) and 510 various texts of children’s literature. The sample contains 372 authors. The age group of potential readers of texts - adult or child - is determined on the basis of expert evaluation (fig. 2).

The minimum length of texts included in the sample is 30 words. More than 60% of texts (623 texts) are shorter than 500 words. The average text length in the sample is 471 words.



	<500	500-1000	1000-1500	1500-2000	2000-2500	2500-3000	3000-3500	>3500
▨ Children's	291	114	32	26	21	17	5	4
■ Adult's	332	118	28	21	13	7	4	0

Fig. 2. The length of the texts in the sample..

The investigation deals with two categories – children’s and adult’s, according to the corpus provided for the experiments.

2.4. Classification features

We have identified a set of features for the texts. These features became the basis for classification feature selecting.

The list of marker features is given in Table 1. The columns "Adult’s" and "Children's" contain the ratio of training sample texts belonging to the category to the total number of training sample texts that have the given feature value.

The values stated in the columns "Adult’s" and "Children's" were produced on the basis of the analysis of the training sample. These values are caused by the corpus. According to the analysis of the training sample, the value "Fairy tale" (the feature “Genre”) is a declarative of belonging to the children’s category. But in practice, the value "Fairy tale" may mark an adult’s text (for example, the books of J. K. Rowling, J. R. R. Tolkien).

Table 1. The list of marker features

Feature	Feature type	Value	Adult’s	Children’s
The presence of profanity	Binary	+	1	0
The presence of special vocabulary	Binary	+	0.95	0.05
The presence of special characters	Binary	+	0.96	0.04
Genre	Nominal	Memoir	1	0
		Fairy tale	0	1
Illustration type	Nominal	Chart	0.98	0.02

The list of interval features is given in Table 2.

Table 2. The list of interval features

Feature	Adult’s	Children’s
Average length of words of text (except stop words)	8.35	6.11
Average number of words in the sentence	11.41	6.2
Percentage of polysyllabic words in the text (more ;yllables)	22.95	13.91
Percentage of special verbal forms in the text	3	2.09
Average number of grammatical foundations of the proposal	2.47	1.81
Percentage of the numerals in the text	3.1	2.59
Percentage of simple sentences with the two main members (among all simple sentences)	64.5	67.3
Percentage of function words	27.85	23.03
Percentage of verbs in the text	20.18	21.79
Percentage of adjectives in the text	11.24	10.83

3. Methods

The original text sample was divided into training and control sample n times. Further, we applied to these texts step-by-step classification algorithm. Firstly, we estimated values of marker features. Then, we excluded categories which do not include the text and made a comparison of the text representing as a set of interval features of the remaining categories. We used two classification methods: evaluation of the distance using the Mahalanobis distance and neural network method.

The formula used to calculate the Mahalanobis distance is:

$$\rho\left(F_{T_i}, R\right) = \sqrt{(F_{T_i} - R)^T \Lambda^T C^{-1} (F_{T_i} - R)},$$
$$R = \frac{\sum_{j=1}^{M} F_{T_j}}{M}, 1 \le M \le L, \tag{1}$$

Λ –matrix of weighting coefficients; *C* –covariance matrix; *R* –vector describing the location of the center of mass of categories; *M* –number of text of the category included in the training sample; *L* – total number of texts.

Neural network method implemented using a simple multi-layer perceptron (fig. 3). Using of this network type is caused by its ability to solve poorly formalized based on existing examples and identifying patterns in the communication of input and output data.

The network architecture was chosen experimentally. The input layer of the neural network comprises a number of neurons equal to the number of classifications, and the output layer contains the number of neurons corresponding to the number of categories. The sigmoid was used as an activation function. The error back propagation algorithm is used for network training.

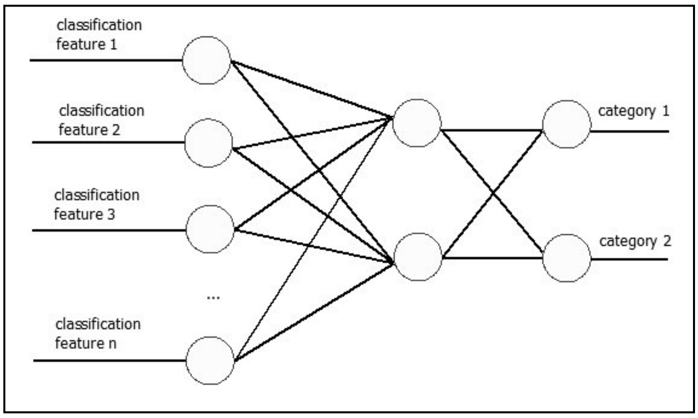


Fig. 3. The network scheme.

Finally, we calculated the average values for all partitions.

4. Results and Discussion

The classification result is a percentage of correctly classified records on the control sample. These values are represented in Table 3. It is obvious that the presented methods show approximately similar results.

Table 3. Results

Method	Accuracy	Standard deviation
Mahalanobis distance	74,16%	5,88%
Neural method	72,07%	6,62%

5. Conclusion

The article describes a step-by-step approach to text classification and results of computational experiment.

We believe that the advantages of this approach are as follows:

1) the normalization of values representing in different feature types to a single range is not required (normalization of interval features is necessary);
2) the time required for system training is reduced;
3) the degree of influence of feature values at the text belonging to the category is evident.

The difficulty of the approach is the need for prior search for marker features.

Acknowledgements

The author wishes to thank all those who participated in the field exercise and helped to make it a successful endeavor.

References

[1] Traum, D. Issues in multiparty dialogues / D. Traum // Advances in Agent Communication. – 2004. – Vol. 1. – P. 201-211.

[2] Jovanovic, N. Addressee identification in face-to-face meetings / N. Jovanovic, R. op den Akker, A. Nijholt Soifer // Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL). – 2006. –P. 249-259.

[3] op den Akker, R. A comparison of addressee detection methods for multiparty conversations / R. op den Akker, D. Traum // Proceedings of DiaHolmia, 13th Workshop on the Semantics and Pragmatics of Dialogue. – 2009. – P. 99-106.

[4] op den Akker, H. Are You Being Addressed? - real-time addressee detection to support remote participants in hybrid meeting / R. op den Akker, H. op den Akker // Proceedings of SIGDIAL. – 2009. – P. 21-28.

[5] Ajvazjan, S. Iterative Applied Statistics: Classification and Reduction of Dimension / S. Ajvazjan, V. Buhshtaber, I. Enjukov, L. Meshalkin – Moscow: Finisy i statistika, 1989. – 607 p.

[6] Database of meta tagging of the Russian National Corpus" (a collection of children's literature). 2014.

[7] Russian National Corpus [Electronic resource]. — Access mode: http://www.ruscorpora.ru/en/ (30.12.2016).