Genre Document Classification Using Flexible Length Phrases

Danijel Radošević, Jasminka Dobša
University of Zagreb, Faculty of Organization and Informatics
Pavlinska 2, 42 000 Varaždin, Croatia
{danijel.radosevic, jasminka.dobsa}@foi.hr

Dunja Mladenić
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Dunja.Mladenic@ijs.si

Zlatko Stapić, Miroslav Novak
University of Zagreb, Faculty of Organization and Informatics
Pavlinska 2, 42 000 Varaždin, Croatia
{zlatko.stapic, miroslav.novak}@foi.hr

Abstract. *In this paper we investigate possibility of using phrases of flexible length in genre classification of textual documents as an extension to classic bag of words document representation where documents are represented using single words as features. The investigation is conducted on collection of articles from document database collected from three different sources representing different genres: newspaper reports, abstracts of scientific articles and legal documents. The investigation includes comparison between classification results obtained by using classic bag of words representation and results obtained by using bag of words extended by flexible length phrases.*

Keywords. Flexible length phrases, bag of words representation, genre classification

1. Introduction

The goal of text categorization is classification of text documents into a fixed number of predefined categories. Document classification is used in many different problem areas involving text documents such as classifying news articles based on their content, or suggesting interesting documents to the web user.

The common way of representing textual documents is by *vector space model* or, so called *bag-of-words representation* [13] . Generally, index term can be any word present in the text of document, but not all the words in the documents have equal importance in representation of document semantic. That is why various schemes in bag-of-words representation give greater weight to words which appear in smaller number of documents, and have greater discrimination power in document classification, and smaller weight to words which are present in lots of documents. Common preprocessing step in document indexing is elimination of, so called, *stop words,* or words such as conjunctions, prepositions and similar, because these words have low discrimination power.

Some approaches extend bag-of-words representation by some additional index terms such as *n*-grams proposed in [11] . Those index terms consist of *n* words forming sequences. In this work, we suggest using *statistical phrases of flexible length.* In the further text we will refer to statistical phrases of flexible length as phrases. The main difference between *n*-grams and phrases, beside the fact that *n*-grams consist of exactly *n* words, opposite to phrases of the flexible length, is that the phrases can contain punctuations. It is important to stress that phrases can also include stop words. The reason for inclusion of stop words is the intention to form phrases characteristic to writing style. So, beside the index terms consisting of just one word that are key words characteristic for some topic, for representation of document we use phrases which could reveal the style of writing. Nevertheless, some phrases are very common, and their classification power is low. That is why we introduce *stop phrases* as phrases that do not appear dominantly in one single category (we used threshold of 70% of all occurrences).

Classification of documents according to style of writing or genre has already been recognized as useful procedure for heterogeneous collections of documents, especially for Web ([4], [7] , [9] ,[10] ,[14]).

Many algorithms are already developed for automatic categorization [6] . For the purpose of our experiments we used the algorithm of support vector machines (SVMs), which was introduced in 1992 by Vapnik and coworkers [2]. Since then, it was shown that algorithm of SVMs is very effective in large scale

of applications, especially for the classification of text documents [8] .
The paper is organized as follows:
- section 1 is introduction to using phrases of flexible length in genre classification of textual documents as an extension to classic bag of words document representation,
- section 2 describes the offered algorithm for generating statistical phrases of flexible length,
- section 3 gives experimental design: the algorithm of the support vector machines (SVM; for classification) and data description
- section 4 gives the results of performed experiments
- section 5 discussion of the experimental results and plans for further work.

2. Generating statistical phrases of flexible length

The algorithm of generating statistical phrases is based on the frequency of the phrases over all the documents. By application of algorithm the list of phrases occurring at least two times in all categories is obtained.

2.1. Algorithm Description

The algorithm [12] starts by dividing all sentences from all the documents into sets of subsentences. Subsentence is a sequence of words that starts from subsequent word position and consist of the rest of the sentence. The minimum length of subsentence is two. Starting from the sentence containing n words, in this way, we get set of (n-1) subsentences. All punctuations are treated in the same way as words.
After that, subsentences are sorted by category. Inside each category they are sorted alphabetically in ascending order. Extracting phrases starts by comparing beginnings of subsequent subsentences. Algorithm for each subsentence finds its overlap with the next subsentence starting form the first word of each subsentence. If overlap consists of minimum of two words it is considered as a phrase.
The algorithm for extracting phrases is summarized as a following pseudocode.

```
Given: Set of documents (each document is a
sequence of sentences consisting of words).

foreach Document
  foreach Sentence
       form a set of  subsentences,
       starting from position of each word in the
       Sentence and taking the rest of the
       sentence
  end // Sentence
end // Document

  collect all subsentences from all documents
  sort subsentences alphabetically
  foreach Subsentence
     compare Subsentence to the next
     subsentence by taking the maximum
     number of overlapping words from the first
     word (forming a phrase)
     extract phrases consisting of minimum two
     words
  end // Subsentence

  foreach Phrase
     count number of occurrences in   documents
  end // Phrase
```

Some of the generated phrases occur frequently in most of the categories and not dominantly in one single category. We consider such phrases as “stop phrases”, i.e. phrases that are useless for the given classification task. Consequently, these phrases are discarded from phrases list.

2.2. Illustration of the algorithm

In the first step all sentences in all documents are divided into sets of all subsentences, starting from each position in the sentence and taking the rest of the sentence (punctuation signs are considered as words; last word is not particularly considered because it can not form a phrase). Illustration of the first step of algorithm is given by the Figure 1.
In the second step all subsentences belonging to same category are put together and sorted.
In the third step each beginning of subsentence is compared to the beginning of next subsentence taking maximum number of overlapping words. Illustration of the second and third step of the algorithm is given by the Figure 2. Extracted phrases are shadowed.

Erozija tla vodom je prirodni proces kojega čovjek može ubrzati.

tla vodom je prirodni proces kojega čovjek može ubrzati.

vodom je prirodni proces kojega čovjek može ubrzati.

je prirodni proces kojega čovjek može ubrzati.

prirodni proces kojega čovjek može ubrzati.

proces kojega čovjek može ubrzati.

kojega čovjek može ubrzati.

čovjek može ubrzat.

čovjek može.

može.

Figure 1: The first step (making subsentences)

. . .

erozija se može svesti na minimum.

erozija tla vodom.

erozija tla vodom je prirodni proces.

erozija zuba kod osoba koje su preboljele karijes.

. . .

nižih količina dušika u proizvodnji.

nižih nadmorskih visina.

nižih ocjena svojim nastavnicima nego

nižih ocjena iz matematike.

nižih prinosa

. . .

Figure 3: The second step (sorting of subsentences) and the third step (extraction of phrases by comparison of beginnings of subsequent sentences). Extracted phrases are shadowed.

3. Experimental design

For classification we used the Support vector machines ([3],[8]), implemented by SvmLight v.5.0 software by Joachims [8] with default parameters. The evaluation is done by 3-fold cross validation.

3.1. The algorithm of support vector machines

Support vector machine ([3],[8]) is an algorithm that finds a hyperplane which separates positive and negative training examples with maximum possible margin. This means that the distance between the hyperplane and the corresponding closest positive and negative examples is maximized. A classifier of the form $sign(w \cdot x + b)$ is learned, where w is the weight vector or normal vector to the hyperplane and b is the bias. The goal of margin maximization is equivalent to the goal of minimization of the norm of the weight vector when the margin is fixed to be of the unit value. Let the training set be the set of pairs $(x_i, y_i), i = 1,2,\ldots,n$ where x_i are vectors of attributes and y_i are labels which take values of 1 and –1. The problem of finding the separating hyperplane is reduced to the optimisation problem of the type

$$\min_{w,b} \quad \langle w, w \rangle$$

$$\text{subject to} \quad y_i\left(\langle w, x_i \rangle + b\right) \geq 1, \quad i = 1,2,\ldots,n.$$

This problem can be transformed into its corresponding dual problem for which efficient algorithms are developed. The dual problem is the problem of the type

$$\max \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle$$

$$\text{subject to} \sum_{i=1}^{n} y_i \alpha_i = 0,$$

$$\alpha_i \geq 0, i = 1,\ldots,n.$$

If $\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*)$ is solution of the dual problem then weight vector $w^* = \sum_{i=1}^{n} y_i \alpha_i^* x_i$ and the bias

$$b^* = -\frac{\max_{y_i=-1}\left(\langle w^*, x_i \rangle\right) + \min_{y_i=1}\left(\langle w^*, x_i \rangle\right)}{2}$$ realize the maximal margin hyperplane. The training pair

(x_i, y_i) for which $\alpha_i^* \neq 0$ is called *support vector*. Only those training pairs influence on calculation of decision function

$$f(x) = w \cdot x + b = \sum_{i=1}^{n} y_i \alpha_i^* \langle x_i, x \rangle + b^*$$

where x is representation of the test document.

3.2. Data description

For the purpose of this investigation we created a collection of Croatian documents consisting of documents from three different sources: 12 263 abstracts of *scientific* papers from Hrvatska znanstvena bibilografija (Croatian scientific bibliography), 2 996 *legal* documents (consisting laws, regulations and similar) from Narodne novine and 2152 *newspaper articles* from Večernji list (all together 17 411 documents). Our hypothesis is that documents from these three sources are written in different style or genre.

A list of features consisting of single words is created based on the criteria that word is contained in at least 10 documents and that word is not contained on the list of stop word. Stop words for Croatian are formed manually as a list of functional words. In such a way we got a list of 13 934 features consisting of single words. Extended list of features (single words + phrases) is formed in a similar way: all words and phrases contained in at least 10 documents are included. Beside the stop words, all stop phrases are discarded. We got a list of 15 220 features consisting of single words and phrases.

4. Experimental results

4.1. List of most frequent phrases

Our algorithm found that in addition to stop phrases, which are as phrases that do not appear dominantly in one single category (we used threshold of 70% of all occurrences), there are some phrases that are typical for only some of categories. In Table 1 we list the ten most frequent phrases from each category.

4.2. Classification results

We performed the classification experiments to test contribution of the flexible length phrases to the classification performance.

For evaluation of classification, we have used the standard measures of recall and precision. Precision p is a proportion of documents predicted positive that are actually positive. Recall r is defined as a proportion of positive documents that are predicted positive.

Table 1: The lists of the most frequent phrases for each category

	Scientific	Legal	News
1.	ovom radu	, 15	rekao je
2.	ovog rada	, 11	ove godine
3.	u ovom radu	Republike Hrvatske je na sjednici	ne može
4.	ovog istraživanja	Vlada Republike Hrvatske je na	i G
5.	u radu se analizira	Hrvatske je na sjednici održanoj	je riječ
6.	u radu je prikazan	, Vlada Republike Hrvfatske je	nije bilo
7.	u razvoju	članka 2	milijuna kuna
8.	Prikaz knjige	stavka 4	dvije godine
9.	i metode	i Republike Hrvatske	oko 1
10.	U radu su prikazani	članka 1	do kraja

Table 2: Precision and recall of document classification for two different representations: bag of words (BOW) and extended representation (BOW+ phrases). In the last row there is macroaverage for all three categories.

	Precision BOW	Precision BOW + phrases	Recall BOW	Recall BOW + phrases
Scientific	99.15 ± 0.35	98.56 ± 0.01	99.75 ± 0.25	99.71 ± 0.16
Legal	99.73 ± 0.12	99.83 ± 0.06	98.80 ± 0.17	98.77 ± 0.12
News	99.45 ± 0.70	99.25 ± 0.44	92.66 ± 1.75	91.96 ± 0.90
Macroaverage	99.44	99.21	97.07	96.81

In Table 2 there are results of classification performance in the measures of precision and recall for two used representations of documents: bag of words (BOW) and extended representation (BOW + phreses). In the last row there is macroaverage of precision and recall computed as a mean of respective measures. The difference between measures of precision and recall obtained by two used representations is tested by pared t-test (p=0.05) for every category separately and it was not found significant difference between these measures obtained by usage of two document representation for any of three categories (scientific, legal and news).

5. Discussion and further work

The experiments have given extremely high results for all categories, and both measures. There no significant differences between results given using bag-of-words representation (13934 features) and extended bag-of-words representation (15220 features). That means that flexible length phrases are useful for document classification, but, unfortunately, don't give better results than standard bag of word representation.

In our future work we plan to investigate classification performances which could be obtained by using significantly reduced number of features including phrases and single words.

6. Acknowledgements

This work was supported by the Slovenian Research Agency and the IST Programme of the European Community under SEKT Semantically Enabled Knowledge Technologies (IST-1-506826-IP) and PASCAL Network of Excellence (IST-2002-506778). This publication only reflects the authors' views. We thank prof.dr. Marko Tadić for list of Croatian stop words.

7. References

[1] R. Baeza-Yates, B.Ribeiro-Neto. *Modern Information Retrieval*, Addison-Wesley, ACM Press, New York, 1999.

[2] B.E. Bosner, I.M. Guyon, V.N. Vapnik. A training algorithm for optimal margin classifier, In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, 1992., pp. 144-152.

[3] N. Cristianini, J. Shave-Taylor. *Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.

[4] A. Dillon, B. Gushrowski, Genres and the Web - is the home page the first digital genre?, *Journal of the American Society for Information Science*, Vol. 51, No.2, pp. 202-205.

[5] J. Dobša, D. Radošević, Z. Stapić, M. Zubac, Automatic categorisation of Croatian web sites, Proceedings of 25th International Convention MIPRO 2005, 2005., 144-149

[6] R.O. Duda, P.E. Hart, D.G. Stork. *Pattern Classification*, second edition, Willey, New York, 2001.

[7] A. Finn, N. Kushmerick, Barry Smyth: Genre Classification and Domain Transfer for Information Filtering, In *Advances in Information Retrieval*, *Proceedings of 24th BCS-IRSG European Colloquium on IR Research*, 2002, pp. 353-362.

[8] T. Joachims. Text categorization with support vector machines: Learning with many relevant features, In *Proceedings of the European Conference on Machine Learning,* 1998, Springer, pp. 137-142.

[9] J. Karlgren, D. R. Cutting. Recognizing Text Genres With Simple Metrics Using Discriminant Analysis, In *Proceedings of 15th International Conference on Computational Linguistics*, 1994, Vol. 2, pp. 1071-1075.

[10] B. Kessler, G. Numberg, N. Schutze. Automatic Detection of Text Genre, In Proceedings of 35th Annual Meeting of Association for Computational Linguistics and 8th Conference of European Chapter of the Association for Computational Linguistics, 1997, pp. 32-38.

[11] D. Mladenić, M. Grobelnik. Word sequences as features in text-learning, In *Proceedings of the 7th Electornical and Computer Science Conference, Ljubljana*, 1998.

[12] D. Radošević, J. Dobša, D. Mladenić: Flexible Length Phrases in Document Classification, In *Proceedings of the 28th International Conference on Information Technology Interfaces*, ITI 2006

[13] G. Salton, C. Buckley. Term-weighting approaches in automatic retrieval, Information Processing & Menagement, 1988, Vol. 24, No. 5, pp. 513-523.

[14] E. Stamatatos, N. Fakotakis, G. Kokkinakis: Automatic Text Categorisation in Terms of Genre and Author , *Computational Linguistics*, 2001, Vol. 26, No. 4, pp. 471-495.