Mining Massive Data Sets for Security
F. Fogelman-Soulié et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-898-4-245

Using linguistic information as features for text categorization

Arturo MONTEJO-RÁEZ [1], Luis Alfonso UREÑA-LÓPEZ,
Miguel Ángel GARCÍA-CUMBRERAS and José Manuel PEREA-ORTEGA
University of Jaén, Spain

Abstract. We report on some experiments using linguistic information as additional features as part of document representation. The use of linguistic features on several information retrieval and text mining tasks is a hot topic, due to the polarity of conclusions encountered by several researchers. In this work, extracted information of every word like the *Part Of Speech*, *stem* and *morphological root* have been combined in different ways for experimenting on a possible improvement in the classification performance and on several algorithms. Our results show that certain gain can be obtained when these varied features are combined in a certain manner, and that these results are independent from the set of classification algorithms applied or the evaluation paradigm chosen, providing certain consistency to our conclusions in text categorization on the Reuters-21578 collection.

Keywords. Automatic text categorization, linguistic features, document representation

Introduction

We report on some experiments using linguistic information as additional features in a classical Vector Space Model [1]. Extracted information of every word like the *Part Of Speech* and *stem*, *morphological root* have been combined in different ways for experimenting on a possible improvement in the classification performance and on several algorithms, like SVM[2], BBR[3] and PLAUM.

The inclusion of certain linguistic features as additional data within the document model is being a subject of debate due to the variety of conclusions reached. This work exposes the behavior of a text categorization system when some of these features are integrated. Our results raise several open issues that should be further studied in order to get more consistent conclusions on the subject. Linguistic features may be useful or not depending on the task, the language domain, or the size of the collection. Nevertheless, we focus here on a very specific aspect: the way we combine features is also crucial for testing its effectiveness.

Automatic Text Classification (TC), or Automatic Text Categorization as it is also known, tries to relate documents to predefined set of classes. Extensive research has been carried out on this subject [4] and a wide range of techniques are applicable to solve this task: feature extraction [5], feature weighting, dimensionality reduction [6], machine

[1]Corresponding Author: Universidad de Jaén, Jaén 23071, Spain; E-mail: amontejo@ujaen.es.

3. Conclusions and future work

Our results show that certain linguistic features improve the categorizer's performance, at least on Reuters-21578. A text classification system shows many degrees of freedom (different tuning parameters), and small variations can produce big deviations, but from the results above, it is clear that for any of the algorithms selected and on any of the evaluation paradigms, the feature combination *word-root-stem-pos* produces better results, but with small improvements compared to the other feature combinations, like morphological root, according to the F1 measure.

Though the gain in precision and recall is not impressive, we believe that further research has to be carried out in this direction, and we plan to study different integration strategies, also considering additional features like *named entities*, term lists and additional combinations of all these features in the aim of finding more synergy. Also, the impact of such information may be higher for full texts than short fragments of Reuters-21578 texts. Collections like the HEP [23] or the JRC-Acquis [24] corpora will be used to analyze this possibility.

At this final point, we would like to underline relevant issues regarding the usage of linguistic features that should also be studied. Some languages (Slavonic languages and Finno-Ugric) are more highly inflected, i.e. there are more variations for the same lemma than, for example, in English. Another important issue is the trade-off between possible errors in the generation of these features by the linguistic tools used and the benefit that their inclusion can produce on the final document representation. Word sense disambiguation may introduce more noise into our data. Also, the stemming algorithm, may perform badly in texts of specialized domains and may harm the final categorization results. Finally, the size of the collection, the length of the document and other characteristics of the data can determine whether the inclusion of certain features is useful or not. Therefore, many questions remain open and the research community still has work to do on this topic.

Acknowledgements

This work has been partially financed by the TIMOM project (TIN2006-15265-C06-03) granted by the Spanish Ministry of Science and Technology and the RFC/PP2006/Id_514 granted by the University of Jaén.

References

[1] Gerard Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. Technical Report TR74-218, Cornell University, Computer Science Department, July 1974.

[2] Thorsten Joachims. Text categorization with support vector machines: learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[3] D. Madigan, A. Genkin, D. D. Lewis, and D. Fradkin. Bayesian multinomial logistic regression for author identification. In *25th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 803 of *AIP Conference Proceedings*. American Institute of Physics, August 2005.