Abstract title

Feature Engineering and Quick Prototyping of Gene Mention Classifiers

Abstract authors

Manuel Maña Lopez (1)
Jacinto Mata Vazquez (1)
Francisco Carrero García (2)
Jose Maria Gomez Hidalgo (2)

Abstract centers-organizations

(1)
Departamento de Ingeniería Electrónica, Sistemas Informáticos y Automática
Escuela Politécnica Superior
Universidad de Huelva
Carretera Huelva - La Rábida
Palos de la Frontera, 21071 Huelva, SPAIN
manuel.mana@diesia.uhu.es, mata@uhu.es

(2)
Departamento de Sistemas Informáticos
Escuela Superior Politécnica
Universidad Europea de Madrid
Villaviciosa de Odon, 28670 Madrid, SPAIN
{francisco.carrero,jmgomez}@uem.es

Abstract

One of the most relevant steps in learning-based Text Classification tasks is the modeling of the task, which is the definition of a suitable set of attributes, amenable of being used by effective learning algorithms. In fact, the learning step is conveniently supported by a number of machine Learning libraries like WEKA and others. Our work is focused on the analysis of the most suitable attributes for a number of Text Classification tasks. We have developed a framework and software library, JTLib, which allows together the analysis, modeling and fast prototyping of classification systems, supporting both the experimentation phase and the development of functional system prototypes. The library provides the essentials of Text Classification currently not provided by WEKA, and in fact, it is a complement to it.

This library is being used in two R&D projects, Isis and Sinamed [1], whose objective is to enhance Information Access in the medical domain through the improvement and utilization of Text Classification tasks, like Text Categorization, Automated Text Summarization, and Biological Entity Recognition.

In the Gene Mention Task we have used our JTLib library and the WEKA package within the following stages:

1. Text indexing. We used JTLib to develop an application that processes the training data (the 15,000 sentences preceded by a identifier) to obtain a representation based on the selected attributes and configured into the input WEKA format (ARFF).
2. Dimensionality reduction. Once the former ARFF file is generated, we used WEKA to process it aiming to find the attributes with best information gain. Then, we obtained a new and definitive training file to build the classifier. We used 28 attributes to characterize each instance.
3. Classifier learning. Using WEKA, we generated a set of models with different Machine Learning algorithms. From these classifiers, C4.5 decision tree achieved the best results. The C4.5 algorithm allows to done a pruned tree in a reduced time but increasing the error rate. We

built two classifiers, both pruned and unpruned.

4. Evaluation of text classifiers. The C4.5 unpruned achieves a scarcely improvement of the F-measure respect to C4.5 pruned. However, the time needed to build the model of the pruned version is a 22% of the time required by the unpruned version. The classification time of the pruned algorithm is also very lower, being the 6% of the time employed by the C4.5 unpruned.

A major (self-imposed) limitation of our approach is that we have not made use of specific resources, apart from the training test itself.

The participation of our team in the Biocreative competition has primary served us as a proof-of-concept for our systematic approach to feature engineering in text classification tasks. We believe we have obtained reasonable results with respect to the effort we have invested in the competitions.

References

[1] Buenaga, M.; Maña, M.; Gachet, D. and Mata, J. *The SINAMED and ISIS Projects: Applying Text Mining Techniques to Improve Access to a Medical Digital Library*. 10th European Conference, ECDL 2006, Alicante, Spain, September 17-22, 2006. 548-551.