Bitext Deep Linguistic Analysis Platform

How linguistic knowledge improves and accelerates training of Machine Learning engines – A Benchmark

The use of POS-tagging information already started this trend years ago. Now, Bitext wants to take this trend to the next level by providing multilingual and multiregister linguistic information. Feature engineering is one of the most critical and time-consuming tasks in Machine Learning for Text Analytics: this is the bottleneck that Bitext solves.

In order to prove that, we have designed a benchmark that compares the performance of a ML engine without and with the linguistic information provided by Bitext Platform. To make the benchmark easy to reproduce, we have taken publicly available datasets as well as open-source ML engines.

Task. As a task for the benchmark, we have chosen text classification, which is the most common application of Machine Learning for text analysis. In particular, we chose a very common use case so it’s easy to grasp and discuss: sentence-level sentiment analysis in English, which involves classifying sentences according their sentiment polarity (positive vs. negative).

Data set. We used the data provided by Kotzias et. all in their paper *From Group to Individual Labels using Deep Features*[1]. The data set consists of labelled sentences extracted from previous data sets covering reviews from IMDB, Amazon and Yelp. For each source, there are 500 positive sentences and 500 negative sentences. As it is customary, we split the data set into 80% for the training set and 20% for the testing set.

Classical engine. Our baseline system employs a "classical" pipeline using standard technologies: an SVM (Suport Vector Machine) trained using bag-of-words (BoW) feature vectors. We used NLTK[2] to generate the BoW features and trained the SVM using svm-light[3]. This pipeline produced a 75% F-score.

[1] https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences

[2] Natural Language Toolkit for Python, http://www.nltk.org/

Engine enhanced with Linguistic Knowledge. Our augmented system was built by adding linguistic knowledge (POS tags and parse trees) to the classical pipeline: an SVM trained with both parse trees and POS-tagged BoW features. We enriched the input used for the baseline system with the parse trees returned by Bitext's Deep Linguistic Analysis platform for English, and trained the system using svm-light-TK[4], an extension for the standard svm-light which is used to support tree kernels (to allow computing the similarity between trees), from Alessandro Moschitti. This pipeline reached over 85% f-score.

Conclusion. The augmented system reached an accuracy of over 85%, a +10% increase over the baseline of the classical engine with an F-score of 75%.

These results are extremely encouraging for the use of linguistic knowledge to boost the performance of ML engines. As next steps, we plan to develop further benchmarks in different lines:

- More complex tasks like ABSA, or topic detection + topic categorization + sentiment
- Other languages
- Other registers or types of texts, particularly informal text

[3] http://svmlight.joachims.org/

[4] http://disi.unitn.it/moschitti/Tree-Kernel.htm