Multi-label text classification using parsimonious language models

Sicco N.A. van Sas SICCO@DDO.NL
University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

Maarten Marx MAARTENMARX@UVA.NL
ILPS, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

Keywords: text classification, information retrieval, parsimonious language models

More and more text documents are available digitally and a need exists to categorize them. Manual indexing is laborious and requires experts, thus there is potential for (semi-)automatic classification of these documents using a controlled vocabulary of concepts. We compare two methods which train classifiers for these concepts. The first method, called JEX, is based the vector space model and was developed by the European Commissions Joint Research Centre (Pouliquen et al., 2003; Steinberger et al., 2012). The second approach, described in this paper, is based on parsimonious language models (PLMs) (Hiemstra et al., 2004) and uses no language-dependent resources. Its parameters are easier to optimize and it outperforms the first approach.

JEX' method is empirically constructed using more than 1500 experiments in which different combinations of formulas and parameters were evaluated. The result is a combination of log-likelihood, which is used to find relevant terms, and a variation of inverse document frequency to calculate the term-weights. Significant improvements are obtained when large (multi-word) stop lists are used and its 9 parameters are fine-tuned.

The PLM method estimates a concept classifier by comparing the language used in documents labeled with that concept to the language used in the whole corpus. Terms which are well enough *explained* in the whole corpus are given a probability of zero, thereby reducing the size of the model and acting as an automatic stop list.

Both methods were trained and compared on two political datasets, Acquis and Dutch parliamentary questions (PQ). We used 19 languages from the Acquis dataset with each between 20.000-42.000 European legislation documents labeled with concepts from the EuroVoc taxonomy, which consist of 6797 hierarchically structured concepts. The PQ dataset contains nearly 40.000 documents labeled with a smaller taxonomy of 111 concepts. The first chronological 90% of the data is used as train set, the final 10% as test set.

Versions of the PLM system based on unigrams and bigrams where found to perform well, as shown in Table 1. The multi-label classifiers yield a ranked list of concepts for each document and are evaluated using R-precision (Rprec) and mean average precision (MAP). The unigram PLM system significantly outperformed JEX in 11/19 languages on the Acquis dataset (and performed similar in 3 other languages), while the bigram version does this for all 19 languages. Both unigram and bigram systems reach significantly higher scores on the PQ dataset. The PLM system uses only 3 parameters though the results across different datasets and languages were obtained using parameters optimized on a single language.

Table 1. Scores for the Dutch Acquis and PQ datasets. Significance tested with two-tailed paired t-tests $^{\blacktriangle} = p < 0.01$.

	JEX (baseline)		**unigram PLM**		**bigram PLM**	
dataset	**Rprec**	**MAP**	**Rprec**	**MAP**	**Rprec**	**MAP**
Acquis (NL)	0.5527	0.5770	0.5576	0.5762	0.5673$^{\blacktriangle}$	0.5906$^{\blacktriangle}$
PQ	0.4120	0.5491	0.4807$^{\blacktriangle}$	0.6197$^{\blacktriangle}$	0.5175$^{\blacktriangle}$	0.6436$^{\blacktriangle}$

References

Hiemstra, D., Robertson, S., & Zaragoza, H. (2004). Parsimonious language models for information retrieval. *Proc. SIGIR'04* (pp. 39–46).

Pouliquen, B., Steinberger, R., & Ignat, C. (2003). Automatic annotation of multilingual text collections with a conceptual thesaurus. *Proc. EUROLAN'03* (pp. 9–28).

Steinberger, R., Ebrahim, M., & Turchi, M. (2012). JRC Eurovoc Indexer JEX - a freely available multi-label categorisation tool. *Proc. LREC'12.*