INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

Data-Based Knowledge Acquisition

Authorship Attribution

In this project, the aim is to build a classifier able to efficiently attribute a new text to his/her author, relying on textual characteristics (words, stems, ngrams of words or characters, punctuation maks, etc.). At least two different representations of the textual data have to be compared.

In order to develop and test the chosen machine learning method, a dataset is provided under Moodle (in the Data directory) (also available at http://archive.ics.uci.edu/ml/datasets.html): The Reuter_50_50 Data Set in which the training corpus consists of 2,500 texts of 50 different authors (50 texts per author) and the test corpus includes other 2,500 texts of the same authors (again 50 per author) non-overlapping with the training texts.

Two main steps have thus to be tackled:

1- Choice of a machine learning method and a first representation of the documents: the aim is to learn a first classifier relying on the representation and to propose a precise evaluation of its performance on the test corpus.

2- Choice of a second type of representation and comparison: relying on new representations of the documents, a new classifier is learnt and its performance is compared to that of the first one.

If useful, a non exhaustive list of natural language processing (NLP) resources and tools is provided, which may be useful (especially if you choose the second version of the project, but not only):

- a lot of segmenters or tokenizers are available; so are stopword lists for several languages (e.g., at http://torvald.aksis.uib.no/corpora/1999-1/0042.html, or Jean Véronis's list but several others exist);
- stemmers or morphological analyzers: Lovins's, Porter's or Paice-Huster's stemmers; Flemm (morphological analyzer for French; F. Namer's website);
- part-of-speech taggers, with or without lemmatization: TreeTagger, Brill, etc.;
- synonyms or paradigmatically related lexical units: WordNet (univ. Princeton or Java version at source.net/projects/jwordnet); the Roget's thesaurus, GREYC's dictionary of synonyms (Caen), etc.;
- corpus-based paradigmatic relation acquisition, using non supervised machine learning techniques; corpus-based semantic relation extraction using ILP, etc.;
- complex term extraction: from French or English textual data, Acabit (B. Daille LINA Nantes), Ana (C. Enguehard LINA Nantes), Lexter (D. Bourigault ERSS Toulouse) and a more extended version Syntex.