INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

Data-Based Knowledge Acquisition

Text Categorization

The aim of this project is

- first to implement and evaluate the performance of a text categorization algorithm, using "standard" representations of the documents;
- second,
  - either to propose a comparative evaluation of the performance of this first categorization algorithm with that of a second one of a different type,
  - or to enrich the representations of the documents with pre-existing or corpus-based linguistic knowledge, and to evaluate the influence of that knowledge on the performance of the algorithm.

Two datasets are provided under Moodle (in the Data directory) to evaluate the categorization algorithms (both also available at http://archive.ics.uci.edu/ml/datasets.html):

- the Reuters-21578 Text Categorization Collection: a collection of 21,578 news items from the news agency Reuters that can be classified into 135 topics (together with the ground truth);
- the Twenty Newsgroups Data Set: 20,000 messages taken from 20 newsgroups (1,000 per newsgroup).

Moreover, 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.

Three main steps have thus to be tackled:

1- Choice of a dataset, and data pretreatment: choose the dataset you want to use in your tests, and format the data in a relevant way wrt the tests. Carefully read the README file for the Reuters's corpus, and the data description on UCI's website for the 20 Newsgroups one.

2- Implementation and evaluation of one categorization algorithm: for this first phase, documents are represented with (some) simple terms. Attention has to be paid to the evaluation methodology.

3- Choice between the two comparison alternatives (new algorithm or new linguistically enriched representations with the same algorithm) **and performance comparison**. If the second alternative is chosen, stemmers can be used; complex terms can serve in the representations, or semantic relations between terms can be exploited. This linguistic knowledge can arise from existing resources or be learnt from your data.