contents

R&D AND TECHNOLOGY TRANSFER

ERCIM News No.50, July 2002

Extending Thematic Lexical Resources by Term Categorization

by Alberto Lavelli, Bernardo Magnini, and Fabrizio Sebastiani

Researchers from IEI-CNR, Pisa, and ITC-irst, Trento, are currently working on the automated construction of specialized lexicons1, as part of an ongoing collaboration in the fields of Machine Learning and Information Retrieval.

Increasing attention is being given to the generation of thematic lexicons1 (ie sets of specialized terms, pertaining to a given theme or discipline). Such lexicons1 are useful in a variety of tasks in the natural language processing and information access fields, including supporting information retrieval applications in the context of thematic, 'vertical' portals.

Unfortunately, the manual generation of thematic lexicons1 is expensive, since the intervention of lexicographers and domain experts working in collaboration is required. Furthermore, a manual approach does not allow for fast response to rapidly emerging needs.
We have developed a methodology for the automatic generation of thematic lexicons1 by 'term categorization', employing a combination of techniques from information retrieval (IR) and machine learning (ML). We view the generation of such lexicons1 as an iterative process of learning previously unknown associations between terms and themes. The process is iterative in that, for each item in a set of predefined themes, it generates a sequence of lexicons1, bootstrapping from an initial lexicon given as input. Associations between terms and themes are learnt from a sequence of sets of generic documents ('corpora'). In this way, the lexicon can be enlarged as new corpora become available. At any given iteration, the process builds the lexicons1 for all the themes in parallel, and from the same corpus.

The method we propose is inspired by recent work in text categorization, the activity of automatically building programs capable of labelling natural language texts with zero, one, or several thematic categories from a predefined set. The construction of an automatic text classifier requires the availability of a corpus of preclassified documents. A general inductive process (called the learner) automatically builds a classifier for the categories of interest by learning their characteristics from a training set of documents.

While the purpose of text categorization is that of classifying documents represented as vectors in a space of terms, the purpose of term categorization is (dually) that of classifying terms represented as vectors in a space of documents. This means that, as input to the learning device and to the term classifiers that it will eventually build, we use 'bag of documents' representations for terms, dual to the 'bag of terms' representations commonly used in text categorization (and information retrieval). In our task, terms are thus items that may belong, and must thus be assigned, to (zero, one, or several) themes belonging to a predefined set. In other words, starting from a set of 'training' terms, a set of 'test' terms is classified, and the test terms which are deemed to belong to a theme are added to the corresponding lexicon; they can then be used as new training terms in the next iteration of the process.

The novelty of this 'supervised' approach is that there is basically no requirement on the corpora employed in the process. This differs from the classic unsupervised approach in which, in order to generate a thematic lexicon for a given topic, a corpus of documents labelled with respect to that topic is needed. This may be problematic, since labelled texts are often hard to obtain, and labelling them requires expert manpower.

As our learning device we adopt AdaBoost.MH(KR), a more efficient and effective variant of the well-known AdaBoost.MH algorithm, developed at IEI-CNR. Both algorithms are an implementation of boosting, a method for supervised learning which has proven one of the best performers in text categorization so far. Boosting is based on the idea of relying on the collective judgment of a committee of classifiers that are trained sequentially. When training the k-th classifier special emphasis is placed on the correct categorization of those training examples which have proven more difficult to classify (ie have been misclassified more frequently) for the previously trained classifiers.

We have chosen a boosting approach not only because of its state-of-the-art effectiveness, but also because it naturally allows for a form of 'data cleaning' if a lexicographer wants to inspect the classified terms for possible misclassifications. At each iteration, in addition to generating the new lexicon, the algorithm ranks the training terms according to their 'difficulty', ie how successful the classifiers generated in the process have been at correctly recognizing their label. Since the highest ranked terms are the ones with the highest probability of having been misclassified in the previous iteration, lexicographers can remove the misclassified examples by scanning the list downwards, stopping when they want. The process of generating a thematic lexicon then becomes an iteration of generate-and-test steps.

We are currently experimenting this methodology using a benchmark called WordNetDomains[42], an extension of WordNet in which each word has been labelled with one or more from a set of 42 themes (such as eg Zoology, Medicine, Sport) commonly used in dictionaries. We have randomly partitioned each thematic lexicon (ie the set of words labelled by one of the 42 labels) into a training and a test set, so that the purpose of the experiment is to check the ability of the algorithm to extract the terms of the test set from the corpus. As our evaluation measure, we are using F1, a combination of precision and recall, commonly adopted in text categorization. As our 'learning' texts, we are using a sequence of subsets of the Reuters Corpus Volume I. Our preliminary experiments have shown this approach to outperform previous approaches to thematic lexicon expansion based on unsupervised methods.

Please contact:
Fabrizio Sebastiani, IEI-CNR
Tel: +39 050 3152 892
E-mail: fabrizio@iei.pi.cnr.it