Building Thematic Lexical Resources by Term Categorization

Alberto Lavelli & Bernardo Magnini
ITC-irst
Via Sommarive, 18 – Località Povo
38050 Trento, Italy
E-mail: {lavelli,magnini}@itc.it

Fabrizio Sebastiani
Istituto di Elaborazione dell'Informazione
Consiglio Nazionale delle Ricerche
56124 Pisa, Italy
E-mail: fabrizio@iei.pi.cnr.it

ABSTRACT

We discuss the automatic generation of *thematic lexicons* by means of *term categorization*, a novel task employing techniques from information retrieval (IR) and machine learning (ML). Specifically, we view the generation of such lexicons as an iterative process of learning previously unknown associations between terms and *themes* (i.e. disciplines, or fields of activity). The process is iterative, in that it generates, for each c_i in a set $C = \{c_1, \ldots, c_m\}$ of themes, a sequence $L_0^i \subseteq L_1^i \subseteq \ldots \subseteq L_n^i$ of lexicons, bootstrapping from an initial lexicon L_0^i and a set of text corpora $\Theta = \{\theta_0, \ldots, \theta_{n-1}\}$ given as input. The method is inspired by *text categorization*, the discipline concerned with labelling natural language texts with labels from a predefined set of themes, or categories. However, while text categorization deals with documents represented as vectors in a space of terms, term categorization deals (dually) with terms represented as vectors in a space of documents, and labels terms (instead of documents) with themes. As a learning device we adopt *boosting*, since (a) it has demonstrated state-of-the-art effectiveness in a variety of text categorization applications, and (b) it naturally allows for a form of "data cleaning", thereby making the process of generating a thematic lexicon an iteration of generate-and-test steps.

1. TERM CATEGORIZATION

The generation of *thematic lexicons* (i.e. lexicons consisting of specialized terms, all pertaining to a given theme or discipline) is a task of increased applicative interest, since such lexicons are important in a variety of tasks pertaining to natural language processing and information access. Unfortunately, the generation of thematic lexicons is expensive, since it requires the intervention of specialized manpower, i.e. lexicographers and domain experts working together. There is thus a need of cheaper and faster methods for answering application needs than manual lexicon generation.

We propose a methodology for the or automatic generation of thematic lexicons by *term categorization*, a novel task that employs a combination of techniques from information retrieval (IR) and machine learning (ML). We view the generation of such lexicons as an iterative process of learning previously unknown associations between terms and *themes* (i.e. disciplines, or fields of activity, or domains). The process is iterative, in that it generates, for each c_i in a set $C = \{c_1, \ldots, c_m\}$ of predefined themes, a sequence $L_0^i \subseteq L_1^i \subseteq \ldots \subseteq L_n^i$ of lexicons, bootstrapping from a lexicon L_0^i given as input. Associations between terms and themes are learnt from a sequence $\Theta = \{\theta_0, \ldots, \theta_{n-1}\}$ of sets of documents (hereafter called *corpora*); this allows to enlarge the lexicon as new corpora from which to learn become available. At iteration y, the process builds the lexicons $L_{y+1} = \{L_{y+1}^1, \ldots, L_{y+1}^m\}$ for all the themes $C = \{c_1, \ldots, c_m\}$ in parallel, from the same corpus θ_y.

The method we propose is inspired by *text categorization* [5], the activity of automatically building, by means of machine learning techniques, *automatic text classifiers*, i.e. programs capable of labelling natural language texts with (zero, one, or several) thematic categories from a predefined set $C = \{c_1, \ldots, c_m\}$.

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. 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 Γ_y^i of preclassified terms, a new set of terms Γ_{y+1}^i is classified, and the terms in Γ_{y+1}^i which are deemed to belong to c_i are added to L_y^i to yield L_{y+1}^i. The set Γ_y^i is composed of lexicon L_y^i, acting as the set of "positive examples", plus a set of terms known not to belong to c_i, acting as the set of "negative examples". For 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.

The novelty of this *supervised* approach to the generation of thematic lexicons is that there is basically no requirement on the corpora $\Theta = \{\theta_0, \ldots, \theta_{n-1}\}$ to be employed in the process. This differs from the classic *unsupervised* approach (see e.g. [1]), in which in order to generate a thematic lexicon for topic c_i, a corpus of documents *labelled by* c_i is needed. This may be problematic, since labelled data are sometimes hard to obtain, and labelling them requires expert manpower. In our approach, the only requirement on θ_y is that at least some of the terms in each of the lexicons in $L_y = \{L_y^1, \ldots, L_y^m\}$ should occur in it (if none among the terms in a lexicon L_y^j occurs in θ_y, then no new term is added to L_y^j in iteration y).

SIGIR'02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.

# of docs	# of training terms	# of test terms	# of labels per term	min # of docs per term	Precision micro	Recall micro	F_1 micro	Precision macro	Recall macro	F_1 macro
2,689	4,424	2,212	1.96	1	0.542029	0.043408	0.080378	0.584540	0.038108	0.071551
2,689	1,685	842	2.36	5	0.512903	0.079580	0.137782	0.487520	0.078677	0.135489
2,689	1,060	530	2.55	10	0.517544	0.086131	0.147685	0.560876	0.084176	0.146383
16,003	7,975	3,987	1.76	1	0.720165	0.049631	0.092863	0.701141	0.038971	0.073837
16,003	4,132	2,066	2.02	5	0.733491	0.075121	0.136284	0.738505	0.065472	0.120281
16,003	2,970	1,485	2.15	10	0.740260	0.091405	0.162718	0.758044	0.078162	0.141712
67,953	11,313	5,477	1.66	1	0.704251	0.043090	0.081211	0.692819	0.034241	0.065256
67,953	6,829	3,414	1.83	5	0.666667	0.040816	0.076923	0.728300	0.050903	0.095155
67,953	5,335	2,668	1.92	10	0.712406	0.076830	0.138701	0.706678	0.056913	0.105342
67,953	4,521	2,261	1.99	15	0.742574	0.086445	0.154863	0.731530	0.064038	0.117766
67,953	3,317	1,659	2.10	30	0.745455	0.098439	0.173913	**0.785371**	0.075573	0.137878
67,953	2,330	1,166	2.25	60	**0.760417**	**0.117789**	**0.203982**	0.755136	**0.086809**	**0.155718**

Table 1: Preliminary results obtained on the automated lexicon generation task. The three blocks of lines correspond to using one day (08.11.1996), one week (08.11.1996 to 14.11.1996), and one month (01.11.1996 to 30.11.1996) of Reuters Corpus Volume 1 newswires, respectively.

As the learning device we adopt AdaBoost.MH^{KR} [6], a more efficient and more effective variant of AdaBoost.MH^{R} [3]. Both algorithms are an implementation of *boosting*, a method for supervised learning which has proven one of the best performers in text categorization applications so far. Boosting is based on the idea of relying on the collective judgment of a committee of classifiers that are trained sequentially; in training the k-th classifier special emphasis is placed on the correct categorization of the training examples which have proven harder (i.e. have been misclassified more frequently) for the previously trained classifiers.

2. EXPERIMENTS

We are currently experimenting this methodology by using WordNet Domains(42) [2] as the lexical resource. WordNet Domains(42) is an extension of WordNet in which each word has been labelled with one or more from a set of 42 themes (such as e.g. Zoology, Medicine, Sport) commonly used in dictionaries. Each thematic lexicon (i.e. the set of the words labelled by the same theme) is partitioned 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 texts. As the evaluation measure, we are using (both micro- and macro-averaged) F_1. As the texts from which to extract the terms, we are using a sequence $\Theta = \{\theta_0, \ldots, \theta_{n-1}\}$ of subsets of the Reuters Corpus Volume 1(see `http://groups.yahoo.com/group/ReutersCorpora/`). Only the text of the documents is used, and the Reuters categories labelling the documents are not used. We lemmatize the documents and annotate the lemmas with part-of-speech tags using the TreeTagger package [4].

Table 1 reports some results from preliminary experiments involving a single iteration of the process and committees of 500 decision classifiers. Only training and test terms occurring in at least x documents were considered (see Column 5); the experiments reported in the same block of lines differ for the choice of the x parameter.

The first conclusion we can draw from them is that F_1 values are still low; a lot of work is still needed in tuning this approach in order to obtain significant categorization performance. The low values of F_1 are mostly the result of low recall values, while precision tends to be much higher, often well above the .70 mark.

The second conclusion is that results show a constant and definite improvement when higher values of x are used, despite the fact that higher levels of x mean (see Column 4) a higher number of labels per term, i.e. more polysemy. This is not surprising, since when a term occurs e.g. in one document only, this means that only one entry in the vector that represents the term is non-null (i.e. significant). This is in sharp contrast with text categorization, in which the number of non-null entries in the vector representing a document equals the number of distinct terms contained in the document, and is usually at least in the hundreds. This alone might suffice to justify the difference in performance between term categorization and text categorization.

3. REFERENCES

[1] H. Chen, B. R. Schatz, T. D. Ng, J. Martinez, A. Kirchoff, and C. Lin. A parallel computing approach to creating engineering concept spaces for semantic retrieval: the Illinois Digital Library Initiative project. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):771–782, 1996.

[2] B. Magnini and G. Cavaglià. Integrating subject field codes into WordNet. In *Proceedings of LREC-2000, 2nd International Conference on Language Resources and Evaluation*, pages 1413–1418, Athens, GR, 2000.

[3] R. E. Schapire and Y. Singer. BoosTexter: a boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.

[4] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK, 1994.

[5] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[6] F. Sebastiani, A. Sperduti, and N. Valdambrini. An improved boosting algorithm and its application to automated text categorization. In *Proceedings of CIKM-00, 9th ACM International Conference on Information and Knowledge Management*, pages 78–85, McLean, US, 2000. ACM Press, New York, US.