RC22969 (W0311-033) November 5, 2003
Computer Science

IBM Research Report

Automatic Construction of Lexical Resources

Rie Kubota Ando
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Automatic Construction of Lexical Resources

Rie Kubota Ando
IBM T.J. Watson Research Center
19 Skyline Dr., Hawthorne, NY 10532
rie1@us.ibm.com

The recognition of entity mentions plays an important role in information extraction systems. For instance, in order to fill a ‘who’ slot of a template for person arrival events, “scientist” needs be recognized as a mention of a person entity. Whether one takes a rule-based approach or employs machine learning techniques, it is useful to have a gazetteer of the words that strongly indicate the mentions of target entity classes.

The necessity has motivated the study of techniques for automatically collecting lexical items with entity categories – sometimes also called the (semi-)automatic construction of *semantic lexicons*. Typically, a method is assumed to be provided with very small amount of input information about the target classes — a few examples (called *seeds*), e.g., { “car”, “plane”, “ship” }. The method seeks to extract more of these from an unannotated corpus. The setting is similar to supervised classification learning in the sense that labeled examples (i.e., seeds) are given to the method. (If negative examples (e.g., non-vehicle) are also available, this is precisely a classification task. And practically, just a few negative examples are very easy to make.) However, the assumed number of labeled examples is believed to be insufficient for general-purpose learning machines to achieve high performance. The rationale of this seemingly atypical task setting is that, at least for us, the outcome (i.e., resultant gazetteers) is for expediting the process of building information extraction systems for new domains with new vocabulary. In this situation, an automatic lexical item collector will be, practically, more valuable if it can produce better results from a very small amount of prior knowledge.

Consequently, previously proposed methods are, more or less, elaborately specialized to the task (as opposed to, say, applying standard linear classifiers). The trend appears to be *bootstrapping* exploiting strong syntactic cues and carefully designed statistical measures (scoring formulas). For instance, Roark and Charniak (1998) has proposed a bootstrapping method that iteratively grows seed sets by adding new words that co-occur with the current seeds in lists, conjunctions, and appositives. The method proposed by Riloff and Jones (1999) is called *meta-bootstrapping* which repeatedly finds extraction patterns (lexical items within some syntactic constructions such as subject-verb) and extracts words from the found patterns, in turn. Thelen and Riloff (2002) report that their new bootstrapping method significantly outperforms the meta-bootstrapping. Phillips and Riloff (2002) combine three bootstrapping processes, each of which exploits one of the following three syntactic constructions: appositives, compound nouns, and ISA-clauses.

The essence of bootstrapping is the automatic and iterative expansion of labeled examples. The motivation is to compensate for the assumed smallness of ‘true’ labeled data. Its potential hazard is seed ‘infection’ or label ‘contamination’ — the phenomenon that wrongly (automatically) labeled examples enter the seed set and misdirect the succeeding extraction. The risk of infection apparently increases as more iterations are performed. It may be difficult to apply bootstrapping in a large scale.

To tackle the problem of automatically collecting lexical items with classes, we start with formalizing the problem setting with respect to the standard framework of classification learning theory. In this framework, we assume that only one seed is given for each class, and precisely quantify the factors that affect the performance of straightforward approaches that compare vector inner products of word occurrence counts. Based

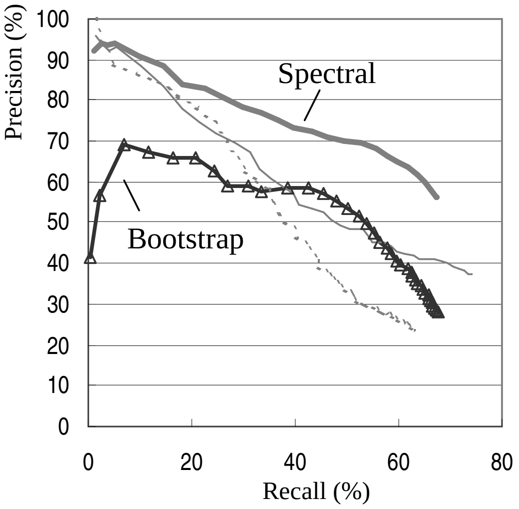

Figure 1: Precision-recall curve.
'Spectral': our method, 'Bootstrap': a bootstrapping method. The other two lines are centroid-based classifiers with and without heuristic feature weighting. The input seeds are 41 in-class words and 59 out-of-class words.

on this analysis, we propose a new method, which employs the *spectral analysis* for approximating 'weighted' conditional probabilities. The method generates feature vectors for use in standard linear classifiers.

We use the term 'spectral analysis' as a more general expression for *singular value decomposition (SVD)*. In the text analysis field, *Latent Semantic Indexing (LSI)* (Deerwester et al., 1990) (applying SVD to term-document matrices for information retrieval) was proposed, and it was followed by a number of empirical studies of its variations. Later, formal analyses of LSI such as (Papadimitriou et al., 1998; Azar et al., 2001; Ando and Lee, 2001), employing the *invariant subspace* perturbation theorems (Davis and Kahan, 1970), appeared. Our work was inspired by these formal analyses; in particular, the theoretical basis of our use of SVD partly derives from our previous work (Ando and Lee, 2001).

Our evaluation was on five target classes – persons, organizations, geo-political entities (GPE), locational entities, and facilities. The definitions of the classes follow the annotation guidelines of ACE (Automatic Content Extraction)[1]. In one setting, the method was tested on a list of approx. 10,000 non-proper nouns. For the purpose of evaluation, these words were manually labeled with six classes – five target classes and 'Others'. The number of words in the target classes were approximately 2,000. Thus, the objective was to extract only those 2,000 words with correct labels (out of 5 classes) from 10,000 words. Observe the difficulty of this task — the chance performance is very low. Figure 1 shows the overall precision-recall curve in the setting where the input seeds are 41 words of the five target classes and 59 'Others'. All the methods use exactly the same syntactic features (such as subject-verb constructions), which are commonly used in previous studies. Our spectral analysis-based method significantly outperforms the other tested methods including a representative bootstrapping algorithm.

A series of experiments strongly support our analysis. In comparison with a representative bootstrapping method, we confirm the following advantages of our approach. First, our method achieves significantly higher F-measure (combined precision and recall), and significantly higher precision at all the inspected recall levels. Second, as noted by several authors, existing approaches often require seeds to be frequently observed in the corpus in order to achieve reasonable performance. In contrast to that, our method is relatively insensitive to the selection of input seeds. This property is particularly beneficial when the items of the target class are sparse

[1] `http://www.nist.gov/speech/index.htm`

in the available corpus. Thirdly, the method can be applied to a very large corpus since its runtime is linear in the size of the corpus.

We will describe our analysis and method in a forthcoming full paper.

References

Rie Kubota Ando and Lillian Lee. 2001. Iterative Residual Rescaling: An analysis and generalization of LSI. In *Proceedings of SIGIR'01*, pages 154–162.

Yossi Azar, Amos Fiat, Anna Karlin, Frank McSherry, and Jared Saia. 2001. Spectral analysis of data. In *Proceedings of STOC 2001*.

Chandler Davis and W. M. Kahan. 1970. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, March.

Scott Deerwester, Susan T. Dumais, Geroge W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the Society for Information Science*, 41:391–407.

C.H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. 1998. Proceedings of symposium on principles of database systems (PODS), seattle, washington.

William Phillips and Ellen Riloff. 2002. Exploiting strong syntactic heuristics and co-training to learn semantic lexicons. In *Proceedings of EMNLP 2002*.

Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*.

Brian Roark and Eugene Charniak. 1998. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In *Proceedings of COLING-ACL'98*.

Michael Thelen and Ellen Riloff. 2002. A bootstrapping method for learning semantic lexicons using extracting pattern contexts. In *Proceedings of EMNLP 2002*.