

Efficient Hierarchical Entity Classification Using Conditional Random Fields

Koen Deschacht Marie-Francine Moens

*Interdisciplinary Centre for Law & IT, K.U.Leuven
Tiensestraat 41, 3000 Leuven, Belgium*

1 Introduction

In this paper we describe work carried out in the CLASS project¹. The central objective of this project is to develop advanced learning methods that allow images, video and associated text to be analyzed and structured automatically. We will, for example, learn the correspondence between faces in an image and persons described in surrounding text. The role of the authors in the CLASS project is mainly on information extraction from text. In the first phase of the project we build a classifier for automatic identification and categorization of entities in texts. This classifier extracts entities from text, and assigns a label to these entities chosen from an inventory of possible labels.

2 WordNet

WordNet [1] is a lexical database that organizes nouns, verbs, adjectives and adverbs in synsets. A synset is a collection of words that have a close meaning and that represent an underlying concept. An example of a synset is “person, individual, somebody, mortal, human being”. WordNet defines a number of relations between synsets. For nouns the most important relation is the hypernym/hyponym relation. A noun X is a hypernym of a noun Y if Y is a subtype or instance of X . For example, “bird” is a hypernym of “penguin” (and “penguin” is a hyponym of “bird”). This relation organizes the synsets in a hierarchical tree. We make the assumption that every synset has exactly one hypernym. This is sensible since only a minority of the synsets has two or more hypernyms. In such a case we choose the most common hypernym. We will build a classifier that tags every noun phrase in a sentence with its WordNet synset using Conditional Random Fields.

3 Conditional Random Fields

Conditional random fields (CRFs) [2] is a statistical method based on undirected graphical models. Let X be a variable over input sequences to be labeled and Y a variable over corresponding label sequences. We define $G = (V, E)$ to be an undirected graph such that there is a node $v \in V$ corresponding to each of the random variables representing an element Y_v of Y . If each random variable Y_v obeys the Markov property with respect to G (e.g., in a first order model the transition probability depends only on the neighboring state), then the model (Y, X) is a Conditional Random Field. Although the structure of the graph G may be arbitrary, we limit the discussion here to graph structures in which the nodes corresponding to elements of Y form a simple first-order Markov chain. CRFs can be thought of as a generalization of both the Maximum Entropy Markov model (MEMM) and the Hidden Markov model (HMM).

When using CRFs to build a classifier for the WordNet synsets, we’ve experienced that the computational complexity of both training and labeling a new sentence is an important delimiting factor when using a very big collection of labels. Employing CRFs to learn the 95076 WordNet synsets was not feasible on current hardware.

¹<http://class.inrialpes.fr/>

4 Reducing computational complexity

We reduce the complexity of both labeling and training using CRFs by taking into account the hierarchical tree of nouns in WordNet. We will first explain how we do this when labeling a new sentence. The standard method to label a sentence with CRFs is the Viterbi algorithm which has a computational complexity of $O(TM^2)$, where T is the length of the sentence and M the number of labels. To reduce this computational complexity we select the best labeling in a number of iterations. In the first iteration, we label every noun phrase in a sentence with a synset chosen from the synsets at the top of the hierarchical tree (“physical entity”, “abstract entity” or “thing”). After choosing the best labeling, we refine our choice (choose a hyponym of the previous chosen synset) in subsequent iterations until we arrive at a synset which has no hyponyms. In every iteration we only have to choose from a very small number of labels, thus breaking down the problem of selecting the correct label from a large number of labels in a number of smaller problems. This results in a computation complexity of $O(T \log_q(M)q^2)$ where q is the average number of hyponyms per synset.

To train the parameters of a CRF we employ the forward-backward algorithm. This algorithm has a computational complexity of $O(TM^2NG)$, where N is the number of training examples and G the number of iterations needed for training. We reduce the complexity of training by making the same assumption as during labeling: at every level of the tree, we select the correct label among the q possible hyponym synsets. All other synsets are left out of consideration. Since the number of levels in the tree is (on average) $\log_q M$, we only have to take into account $q \log_q M$ synsets during training, resulting in a complexity of $O(T(q \log_q M)^2 NG)$.

5 Results

We used the Semcor corpus [1] for training. In this corpus, that contains almost 700,000 words, every sentence is noun phrase chunked. The chunks are tagged with their part-of-speech tag and their WordNet synset. We implemented the described method (by adapting the Mallet² package) and trained for approximately 102 hours. Testing on unseen data resulted in an accuracy of 77.82%. We must note that a baseline approach that ignores context completely but simply assigns the most likely sense (according to WordNet) to a given word already achieves an accuracy of 67%. Furthermore, Mihalcea and Moldovan [3], who use the semantic density between words to determine the word sense, achieved an 86.5% accuracy on the same data.

An important weakness of our algorithm is the fact that, to label a sentence, we have to traverse the hierarchy and choose the correct synset at every level. An error at a certain level can not be recovered. In the future, we plan to improve our method by implementing a beam-search, keeping a number of possible synsets at every level.

Acknowledgments

The work reported in this paper was supported by the EU-IST project CLASS (IST-027978).

References

- [1] C. Fellbaum, J. Grabowski, and S. Landes. Performance and confidence in a semantic annotation task. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.
- [2] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, 2001.
- [3] R. Mihalcea and D.I. Moldovan. A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th conference on Association for Computational Linguistics*, pages 152–158. Association for Computational Linguistics Morristown, NJ, USA, 1999.

²<http://mallet.cs.umass.edu>