Proceedings of HLT-NAACL 2003
Short Papers , pp. 64-66
Edmonton, May-June 2003

Semantic Extraction with Wide-Coverage Lexical Resources

Behrang Mohit
School of Information Management & Systems
University of California, Berkeley
Berkeley, CA 94720, USA
behrangm@sims.berkeley.edu

Srini Narayanan
International Computer Science Institute
Berkeley, CA 94704, USA
snarayan@icsi.berkeley.edu

Abstract

We report on results of combining graphical modeling techniques with Information Extraction resources (Pattern Dictionary and Lexicon) for both frame and semantic role assignment. Our approach demonstrates the use of two human built knowledge bases (WordNet and FrameNet) for the task of semantic extraction.

1. Introduction

Portability and domain independence are critical challenges for Natural Language Processing (NLP) systems. The ongoing development of public knowledge bases such as WordNet, FrameNet, CYC, etc. has the potential to support domain independent solutions to NLP. The task of harnessing the appropriate information from these resources for an application remains significant. This paper reports on the use of semantic resources for a necessary component of scalable NLP systems, *Semantic Extraction (SE)*.

Semantic Extraction pertains to the assignment of semantic bindings to short units of text (usually sentences). The SE problem is quite similar to the Information Extraction (IE) task, in that in both cases we are interested only in certain predicates and their argument bindings and not in full understanding. However there are major differences as well. IE is a pre-specified and autonomous task with a narrow domain of focus, where all the information of interest is represented in the extraction template. SE involves finding predicate-argument structures in open domains and is a crucial semantic parsing step in a text understanding task.

In this paper we report results obtained from combining IE and graphical modeling techniques, with semantic resources (WordNet and FrameNet) for automatic Semantic Extraction.

2. Background

Semantic Extraction has become a strong research focus in the last few years. A good example is the work of Gildea and Jurafsky (2002) (GJ). GJ present a comprehensive empirical approach to the problem of semantic role assignment. Their work looked at the problem of assigning semantic roles to text based on a statistical model of the FrameNet[1] data. In their work, GJ assume that the frame of interest is determined a-priori for every sentence.

In the IE community, there has been an ongoing effort to build systems that can automatically generate required pattern sets as well as the extraction relevant lexicon. Jones and Riloff (JR) (1999) describe a bootstrapping approach to the problem of IE pattern extension. They use a small seed lexicon and pattern set, to iteratively generate new patterns and expand their lexicon until they achieve an optimized set of patterns and lexicon.

In the area of lexicon acquisition, many researchers have employed public knowledge bases such as WordNet in IE systems. Bagga et. al. (1997) and later Harabagiu and Maiorano (HM) (2000) investigated the acquisition of the lexical concept space using WordNet and have applied their methods to the Information Extraction task.

In this paper, we describe work that blends the semantic labeling approach exemplified by the GJ effort and the bootstrapping approach of JR and HM. Our work differs from the previous efforts in the following respects. 1) We used FrameNet annotations as seeds both for patterns and for the extraction lexicon. We expand the seed lexicon using WordNet. 2) We built a graphical model for the semantic extraction task, which allows us to integrate automatic frame assignment as part of the extraction. 3) We employed IE methods (including pattern sets and Named Entity Recognition) as initial extraction steps.

[1] http://www.icsi.berkeley.edu/~framenet

3. FrameNet

FrameNet (Baker et. al. 1998) is building a lexicon based on the theory of Frame Semantics. Frame Semantics suggests that the meanings of lexical items (lexical units (LU)) are best defined with respect to larger conceptual chunks, called Frames. Individual lexical units *evoke* specific frames and establish a binding pattern to specific slots or roles (frame elements (FE)) within the frame. The Berkeley FrameNet project describes the underlying frames for different lexical units, examines sentences related to the frames using a very large corpus, and records (annotates) the ways in which information from the associated frames are expressed in these sentences. The result is a database that contains a set of frames (related through hierarchy and composition), a set of frame elements for each frame, and a set of frame annotated sentences that covers the different patterns of usage for lexical units in the frame.

3.1 FrameNet data as seed patterns for IE:

Using the FrameNet annotated dataset, we compiled a set of IE patterns and also the lexicon for each of the lexical units in FrameNet.

We filtered out all of the non-relevant terms in all frame element lexicons. We hypothesized that using a highly precise set of patterns along with precise lexicon should enable a promising IE performance. For our Information Extraction experiments, we used GATE (Cunningham et. al. 2002), an open source natural language engineering system. The component-based architecture of GATE enabled us to plug-in our FrameNet based lexicon and pattern set and run IE experiments on this system.

3.2 Initial Experiment:

As a preliminary test, we compiled a set of 100 news stories from Yahoo News Service with topics related to Criminal Investigation. We also compiled a set of IE patterns and also the lexicon from the crime related frames ("Arrest", "Detain", "Arraign" and "Verdict".) We ran the GATE system on this corpus with our FrameNet data. We evaluated the IE performance by human judgment and hand counting the semantic role assignments. The systems achieved an average of 55% Recall while the precision was 68.8%. The fairly high precision (given just the FrameNet annotations) is the result of a highly precise lexicon and pattern set, while we see the low recall as the result of the small coverage. That is the reason that employed WordNet to enlarge our lexicon.

4. Expanding the Lexicon

In order to expand our lexicon for each of the frame elements, we used the human-built knowledge base (WordNet (Fellbaum 1998)) and its rich hierarchical structure.

We built a graphical model of WordNet making some assumptions about the structure of the induced WordNet graph. For our initial experiments, we built a graph whose leaf was the enclosing category of the FrameNet annotated frame element. We then looked at an ancestor tree following the WordNet *hypernym* relation. This gave us a graphical model of the form shown in Figure 1 for the FrameNet frame element *Suspect* and WordNet category *Thief.*

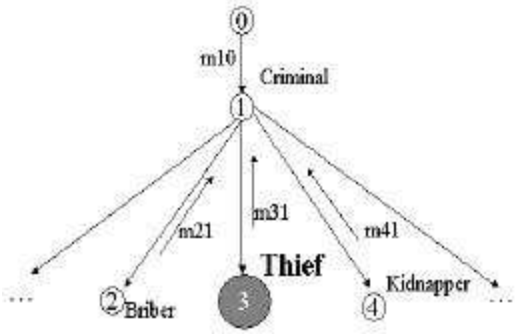

Figure 1

We then used the sum-product algorithm (Frey 1998) for statistical inference on the frame element induced graph (such as in Figure 1). We now illustrate our use of the algorithm to expand the FrameNet derived lexicon.

4.1 Statistical Inference

We employed a statistical inference algorithm to find the relevant nodes of WordNet. For each of the frame elements, we took the terms in FrameNet FE annotations as ground truth which means that the relevance probability of the WordNet nodes for those terms is equal to 1. The Sum Product algorithm helps us find the relevance probability of higher level nodes as a lexical category for the frame element through a bottom up computation of the inter-node messages. For example the message between nodes 1 and 0 in the Figure 1 can be computed as:

$$m_{1,0}(N_0) \propto \sum_{N_1} P(N_0)P(N_1 \mid N_0) \prod_{k \in N(1) \setminus 0} m_{k1}(N_1)$$

We should note that based on the WordNet's hypernym relation, the conditional relevance probability of each parent node (given any child node) is equal to 1. Therefore the Sum Product inter-node messages are computed as:

$$m_{ji}(N_i) = \sum_{N_j} P(N_j) \prod_{k \in N(j) \setminus i} m_{kj}(N_j)$$

and the probability of each WordNet node can be computed by a normalized interpolation of all of the incoming messages from the children nodes:

$$p(N_i) = \frac{\prod_{j \in N(i) \setminus parent(i)} m_{ji}(N_i)}{|N(i) \setminus parent(i)|}$$

4.2 Relevance of a WordNet Nodes

Throughout our experiments with the training data, we discovered that some infrequent tail terms in the frame element lexicon that might not be filtered out by the statistical inference algorithm but still are frequently used in relevant text.

Therefore, we defined the *relevance* metric for the WordNet nodes to achieve a larger coverage. We compiled a large corpus of text (News stories) and made a second smaller corpus from the original one which contains only sentences which are relevant to the IE task. For each of the WordNet nodes we defined the relevance of the node based on the proportion of the occurrence of the node in IE related Text (Orel) to the occurrence of the node in the general text (Ogen).

$$Rel(N) = \frac{O_{rel}}{O_{gen}}$$

Using this relevance metric, we evaluated all of the WordNet nodes for the training data (found in the previous step) and re-ranked and picked the top 'm' relevant nodes (m=5 for our reported experiment) and added them to the previous set of WordNet nodes.

With a set of relevant WordNet nodes, we extended the lexicon for the IE system and re-ran our IE task on the same 100 Yahoo news stories that were used in the initial experiments. The average recall rose up to 76.4% this time with an average precision equal to 66%.

5. Frame Assignment

Using FrameNet data with IE techniques shows promising results for semantic extraction. Our current efforts are geared toward extending the extraction program to include automatic frame assignment. For this task, we assume that that the frame is a latent class variable (whose domain is the set of lexical units) and the frame elements are variables whose domain is the expanded lexicon (FrameNet + WordNet). We assume that the frame elements are conditionally independent from each other, given the frame. For our initial experiments, we assume that each frame is an independent model and frame assignment is the task of selecting the Maximum A Posteriori (MAP) frame given the input and the priors of the frame. Figure 2 shows the graphical model exemplifying this assertion. With this model, we are able to estimate the overall joint distribution for each FrameNet frame, given the lexical items in the candidate sentence from the corpus. During training frame priors and model parameters $p(fe|frame)$ are estimated from a large corpus using our SE machinery outlined in sections 3 and 4. While our initial results seem promising, the work is ongoing and we should have more results to report on this aspect of the work by the time of the conference.

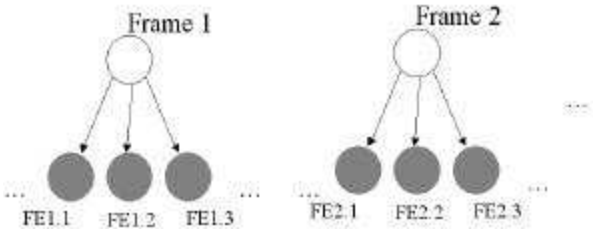

Figure 2

6. References

Bagga A., Chai J.Y. & Biermann A. 1997. The Role of WordNet in The Creation of a Trainable Message Understanding System. In *Proceedings of the Sixth Message Understanding Conference on Artificial Intelligence (AAAI/IAAI-97)*

Baker C., Fillmore C. & Lowe J. 1998, The Berkeley FrameNet project, In *Proceedings of COLING/ACL* pages 86–90, Montreal, Canada.

Cunningham H., Maynard D., Bontcheva K., Tablan V. 2002, GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).*

Fellbaum C., *WordNet: an Electionic Lexical Database,* Cambridge, MA, The MIT Press.

Frey B.J. 1998, *Graphical Models for Machine Learning and Digital Communication*, Cambridge, MA, MIT Press

Gildea D., Jurafsky D. 2002, Automatic labeling of semantic roles, *Computational Linguistics*, 28(3):245-288.

Harabagiu S., Maiorano, S. 2000, Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction, *in Proceedings of LREC-2000,Athens Greece.*

Riloff, E. and Jones, R. 1999, Learning Dictionaries for Information Extraction by Multi-Lev