Semantic Integration in Learning from Text

Steven Bethard, Rodney Nielsen, James H. Martin and Wayne Ward

Department of Computer Science
University of Colorado at Boulder
430 UCB, Boulder, CO 80309, USA
{steven.bethard, rodney.nielsen, james.martin, wayne.ward}@colorado.edu

Abstract

We define learning as the generation of meaningful knowledge representations which can be utilized in future decision making. Optimal learning entails that these knowledge representations be integrated with prior knowledge. In this paper, we introduce a knowledge representation based on an integration of a variety of shallow semantic parsing techniques. Entity detection, event detection, semantic role labeling and temporal relation identification are combined to produce graph-like structures which represent the most important semantic components of a text and the relations between these components. We show how new entities, events and relations can be successfully integrated into this representation using features derived from lexical and dependency-based sources.

Introduction

Advances in machine learning and natural language processing have now put within reach the generation of complete semantic representations from large corpora. These knowledge bases will facilitate improvements in a wide range of tasks, such as question answering, automated tutoring, and multi-document summarization. In question answering, such a properly indexed knowledge base would not only result in much faster and more accurate systems, but would also ease the process of answering questions that would otherwise require extracting and merging information from multiple documents. In automated tutoring, it would be useful in verifying the accuracy of students' answers, discovering prior knowledge that the student seems not to exhibit, generating Socratic dialogs based on the relational information in the knowledge base, and suggesting analogies or related concepts that might facilitate comprehension.

We define learning as the generation of meaningful knowledge representations which can be utilized in future decision making. The key to productive learning from text is not just linking the internal entities and events, but also integrating these relationships with existing knowledge. In work on human text comprehension, Kintsch (1998) calls these two parts the textbase and the situation model, respectively. Examining what we know about human text comprehension might provide some insight into important considerations for machine reading. Given an average paragraph, humans can generate an immense number of implicatures and entailments. However, research has shown that people generate relatively few of these inferences online while reading text (Kintsch 1998). The inferences generated online tend to be mostly those that are required to maintain coherence (e.g., coreference resolution), explain events (e.g., causal antecedents) or support specific reader goals.

Similarly, we suggest that in machine reading it would be a mistake to over-generate inferences at the time of building a representation and rather put it off until the time they are needed. In addition to entity and event coreference resolution, we believe inferences should be drawn that provide important temporal relations between events, causal connections, and where possible, inferences that connect groups of events from a single document that would otherwise remain isolated.

In computational linguistics, some of this semantic integration has been addressed piecewise in prior literature, for example, entity coreference resolution (see Olsson 2004 for an overview) and semantic role parsing (Surdeanu et al. 2003; Pradhan et al. 2005; Toutanova, Haghighi, and Manning 2005). The remaining aspects have seen far less research, (e.g., event coreference resolution, event temporal relation identification, cross-document entity and event coreference resolution, and deriving important implicatures). In this paper, we introduce a knowledge representation which integrates these relations, as output by a variety of shallow semantic parsing techniques.

In the following sections, we first describe our approach to entity and event detection, and then discuss semantic role labeling and temporal relation identification. The outputs of these systems are combined to produce a graph-like structure which represents the most important semantic components of a text and the relations between

these components. Using this semantic representation, we show how new entities, events and relations can be successfully integrated using features derived from lexical and dependency-based sources.

Entity Detection

As our most basic semantic component, we consider the entities in the text, that is, the people, places and things that participate in the various events of the document. Much work has been done on the extraction of such entities, encouraged substantially by competitions like MUC (Grishman and Sundheim 1996) and NIST's Automatic Content Extraction (ACE) task[1]. For our purposes, it is important not only to know where entities are mentioned in the text, but also to know which mentions are referring to the same real-world entity. Thus we are interested in both entity detection and entity coreference.

Our entity system is based on the time and entity mention labelers of (Hacioglu, Chen and Douglas 2005) and (Hacioglu, Douglas and Chen 2005). These labelers follow a word-chunking paradigm and attempt to annotate each word in the text as Beginning, Inside or Outside of an entity or time mention. Combining word-level features like part-of-speech tags and syntactic base-phrase labels with support vector machine (SVM) classifiers, they achieve F-measures in the mid 80s for both of these tasks.

In order to cluster these entity mentions into real-world entities, we follow current state-of-the-art approaches and first train a classifier to identify the likelihood of two entity mentions being coreferential. Then we apply an agglomerative clustering algorithm to these entity mention pairs to group them appropriately[2]. In the end, this produces a simple representation of our text: the set of real-world entities that it discusses.

Event Detection

Knowing which entities are referred to in a text tells us something about that text, but without knowing what events those entities are involved in, we are missing much of the text's meaning. To address this problem, we first identify the words in the document that indicate which events are taking place. This might seem like a simple task – just label all verbs as events and be done with it. However, events don't always appear as verbs, e.g. *the destruction of the city*, and all verbs don't appear as events, e.g. support verbs like *make* in *make a decision*.

Recent work has made some progress in this area however, and we adopt the model of (Bethard and Martin 2006) to locate events in our texts. This approach treats event detection as a word-chunking task and uses word features like affixes, part-of-speech tags and hypernyms in WordNet to train a classifier that can distinguish events from non-events with F-measures in the 70s and 80s. After applying this model to our text, we can now represent it as both a set of real-world entities, and a set of events.

Relation Detection

Of course, true understanding of the text requires more than just knowing what entities and events are involved. Truly understanding the text means recognizing that these entities and events are tied together in various ways. It is these relations between entities and events that are at the core of text understanding, and thus at the core of our knowledge representation. Currently, we are considering two main systems for extracting such relations: semantic role labeling and temporal relation labeling.

Semantic Role Labeling. There has been a flurry of recent research on semantic role labeling, a task in which models are trained to identify the arguments of a predicate (Surdeanu et al. 2003; Pradhan et al. 2005; Toutanova, Haghighi, and Manning 2005). These models can associate a predicate with the phrases it relates, so that for a predicate like *give* in the sentence *John gave his sister the book*, these systems can identify *John* as the Agent, *his sister* as the Beneficiary and *book* as the Theme. Typically, this research has focused on the arguments of verbal predicates, though recent research has shown some success on nominal predicates as well (Jiang and Ng 2006).

We use the ASSERT system of (Pradhan et al. 2005), which uses support vector machine classifiers to inspect each phrase in a syntactic tree and determine whether or not that phrase is an argument of the given predicate. By employing a variety of syntactically informative features, ASSERT is able to find and label the predicate argument phrases with F-measures in the mid-80s.

Temporal Relation Labeling. While many semantic role labelers (ASSERT included) produce Temporal roles, they seldom distinguish between the different types of temporal relations. However, for text understanding, it is crucial to know, for example, which of *Hezbollah fired rockets* or *Israel launched airstrikes* came first.

(Mani, et. al. 2006) made some finer-grained distinctions here, classifying temporal relations as one of the following types: Before, ImmediatelyBefore, Begins, Ends, Includes and Simultaneous. They showed that, given a temporal relation of an unknown type, they could identify the appropriate label over 90% of the time using a maximum entropy model and features like the tense, aspect and modality of the predicates. Thus, these temporal relations, which are so crucial for understanding textual timelines, are now within reach of our current statistical methods.

[1] http://www.nist.gov/speech/tests/ace/

[2] See http://sds.colorado.edu/EXERT/ for more details about this approach and an online demo

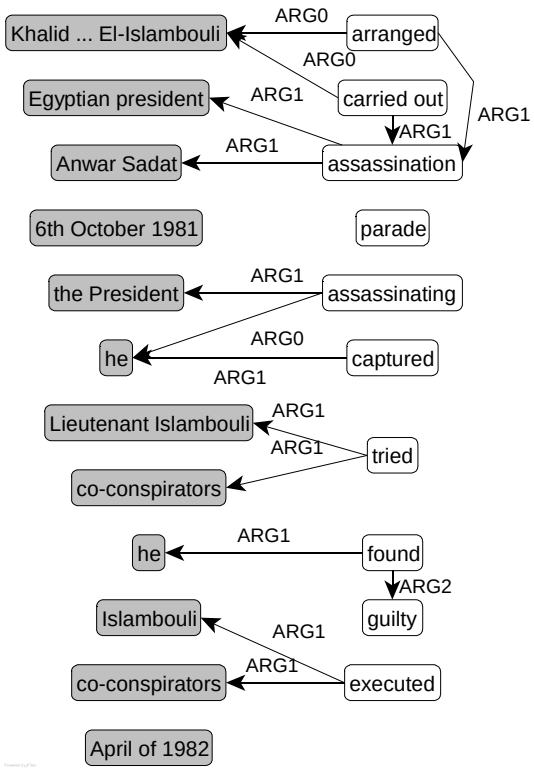

Figure 1: Semantic graph of entities, events and semantic roles.

Graph Generation

Thus we can see a variety of elements of meaning that can now be automatically extracted from text: entities, events, semantic relations and temporal relations. To convert these surface-level semantic descriptions of a text into a deeper-level representation, we assemble them together into a semantic graph.

Events and entities form the nodes of the graph, with the edges between these nodes derived from the semantic and temporal relations. Since semantic roles are defined as phrases, not individual entities or events, we convert these event-phrase relations into event-entity or event-event relations by linking the event to the semantic head-word[3] of the phrase. So in a sentence like *Over three hundred Islamic radicals were indicted*, where the predicate *indicted* has the Theme *over three hundred Islamic radicals*, we identify the relation Theme(*indicted, radicals*). In essence, this process converts our semantic roles to semantic word dependencies. This conversion is crucial to our approach as our graphs describe links between real-world entities and events, not between words and phrases. Figure 1 shows the results of applying such a process to the following text:

> Khalid Ahmed Showky El-Islambouli arranged and carried out the assassination of the Egyptian president, Anwar Sadat, during the annual "6th October 1973 victory" parade on 6th October, 1981. Immediately after assassinating the President, he was captured. Lieutenant Islambouli and twenty-three co-conspirators were tried, and he was found guilty. Islambouli and five others co-conspirators were executed in April of 1982.

Semantic roles play a large part in linking together entities and events into one cohesive knowledge representation, but since semantic roles do not cross sentence boundaries, relying on them alone means having a disconnected graph where at best there is one component for each sentence. (You can see a more realistic case in Figure 1 where three sentences have produced a graph with ten components.) Hence, we rely on two other associative forces to build more fully connected representations of the text: entity coreference and temporal relations.

When two mentions of an entity are known to be coreferential, we merge their nodes into one. By consolidating our knowledge in this way, we gain the ability to reason over all the relations in which a single entity participates. For example, in the graphs for the sentences *Khalid Islambouli carried out the assassination of Anwar Sadat* and *Immediately after assassinating the President, he was captured*, the entity nodes for *Islambouli* and *he* are coreferential. After merging them, we can conclude that *Islambouli* was a participant in both an *assassination* event and a *captured* event, and while he played an Agent role in the former, he played a Theme role in the latter. By merging coreferential nodes, we form a more cohesive knowledge base for the entities in the text.

The other source of cross-sentence relations we rely on are the temporal relations of (Mani et. al. 2006). Since these are defined directly as event-event or event-time relations, with no reliance on phrases or syntactic trees, they quite frequently tie together nodes from different sentences. In the TimeBank corpus (Pustejovsky et. al. 2003), on which the Mani et. al. work is based, around 75% of Before relations[4] are between events or times in different sentences. Consider an example like:

> Lieutenant Islambouli and twenty-three co-conspirators were tried, and he was found guilty.

[3] We use a set of tree-walking rules much like the syntactic head-word rules used by many lexicalized syntactic parsers, but with the rules modified to prefer nouns, verbs and adjectives as heads rather than the prepositions and complementizers.

[4] After temporal closure is performed – see (Mani et. al. 2006) for details.

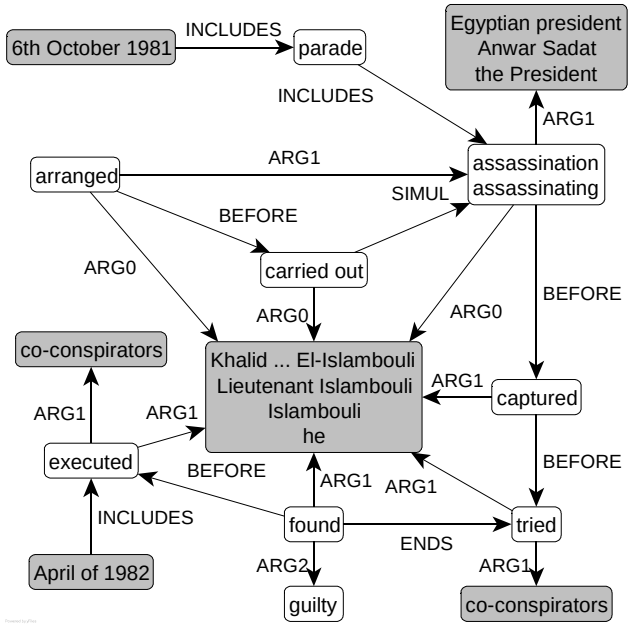

Figure 2: Semantic graph from Figure 1 after adding coreference and temporal relations. Nodes referring to Khalid Islambouli have been merged into one, as have nodes referring to Anwar Sadat. Temporal relations have been added that link the events in chains like *captured → tried → found guilty → executed*. Note that in this graph, temporal relations that were inferable from existing ones were omitted for the sake of clarity.

> Islambouli and five others co-conspirators were executed in April of 1982.

Identifying the Before relation between *found guilty* and *executed* ties these two sentences more tightly together in our representation and in combination with the other temporal relations allows us to perform simple temporal reasoning to conclude that, for example, the co-conspirators were tried *before* Islambouli was executed. Figure 2 shows the result of adding such temporal links (along with entity coreference) to the graph of Figure 1.

In general then, we see that the integration of semantic roles, entity coreference and temporal relations produces a connected, cohesive knowledge representation that can be used to identify the important entities and events in a text, and reason about the relations between them.

Integrating New Knowledge

Thus far we have integrated entities, events, semantic roles and temporal relations within a single document. With this integration complete, we turn our attention to integrating information across documents. Cross-document integration can be used to reinforce the confidence of existing relationships and to insert new supporting and elaborative facts that are tied to the existing entities and events.

The first step in integrating a new document into the knowledge representation is identifying which entities and events of the new document are potentially referring to entities and events in the existing semantic graph. To be able to search for such entities and events, our knowledge representation must be indexed in such a way that subsets of existing nodes and relations that are similar to a new document can be easily retrieved. There has been some work in indexing such nodes and relations, in particular, the Carnegie Mellon JAVELIN question answering system built indexes of semantic roles in addition to the usual term-based indexes (Nyberg et. al. 2005). When retrieving the results for a query, JAVELIN consulted both the term-based index and the role-based index in order to select the most appropriate documents. An approach like this means that we can apply information retrieval techniques to automatically select candidate entities and events for integration with the new information.

So, given a set of potentially co-referring entities and events, we use a machine learning classification approach based on (Nielsen, Ward and Martin 2006) to determine whether the entities and events referred to are in fact the same. This approach determines whether one relation between entities or events is a paraphrase of another based on a set of lexical, dependency, and dependency path similarity features. We briefly sketch these features in the following paragraphs.

We generate a set of lexical similarity features based loosely on the pointwise mutual information for term-document co-occurrence and distributional similarity statistics. These features help identify similar relations anchored to similar terms (e.g. *Islambouli was tried* and *the trial of Islambouli*) while ruling out integration of a new relation when some semantic arguments are unrelated at the lexical level (e.g. *Islambouli was captured* vs. *Islambouli was executed*).

Lexical similarity can be deceiving, however. Consider the task of comparing a child's knowledge of physics with an existing physics knowledge base. The existing knowledge base would contain information like *vibrations are movements and vibrations produce sound*. When a child produces a sentence like *sounds vibrate and hit the object and it moves*, we must recognize that this is in conflict with our knowledge base – the child has confused cause and effect. However, at the lexical level these two sentences match almost perfectly.

To address this issue we generate similarity features based on comparisons between the relations in our semantic graph and the relations in dependency parses. Using statistical corpus information, we can identify patterns like the high mutual information between the dependency pattern Mod(PERSON, *assassin*) and the role

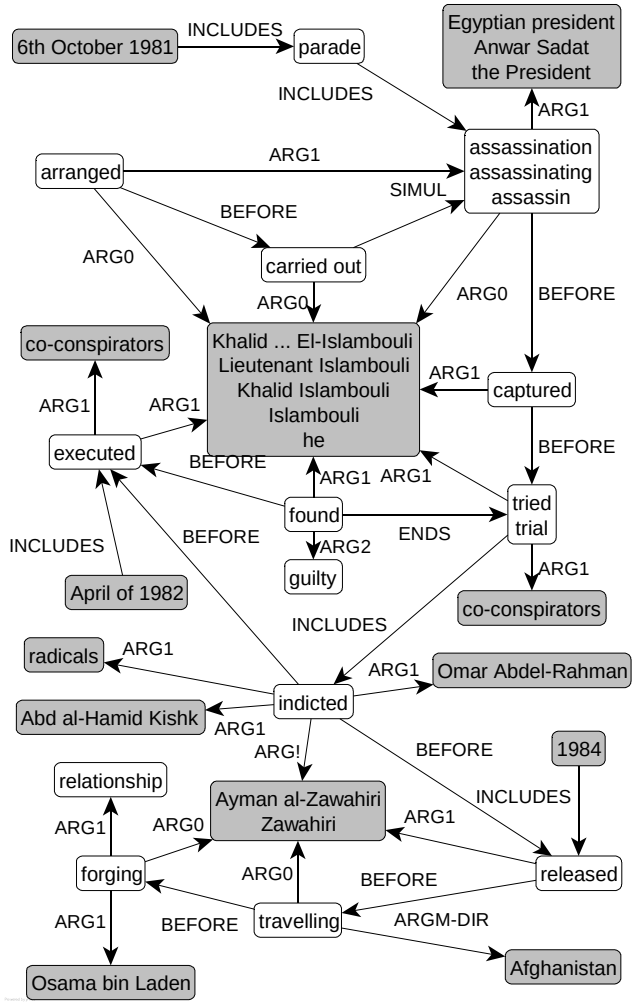

Figure 3: Integration of new text with the semantic graph of Figure 2. The new phrase *trial of assassin Khalid Islambouli* has been matched to the existing relations ARG0(*assassination, Islambouli*) and ARG1(*tried, Islambouli*), drawing with it the new related information. Again, some temporal links have been omitted for clarity.

pattern Agent(*assassinate*, PERSON). When we consider a text like *radicals were indicted in the trial of assassin Khalid Islambouli*, these mutual information features come into play and suggest that Mod(*Islambouli*, *assassin*) is a good candidate for integration with a relation like Agent(*assassinate*, *Islambouli*).

We also generate features based on distinct paths between coreferring entities following (Lin and Pantel 2001). For example, examining a corpus and determining that there is a high mutual information between the path PERSON1 ← *carried out* → *assassination* → PERSON2 and the path PERSON1 ← *was captured* → *after* → *assassinating* → PERSON2, we can infer that the following two sentences are likely partial paraphrases of each other: *Islambouli carried out the assassination of Anwar Sadat* and *After assassinating Sadat, Islambouli was captured*. These dependency-path features are particularly important as they can account for a wide range of paraphrases, while still restricting the relations to those that actually exist in real text.

Figure 3 shows the result of applying these lexical and dependency features to integrate the following sentences with our knowledge base of Figure 2.

> Over three hundred Islamic radicals were indicted in the trial of assassin Khalid Islambouli, including Ayman al-Zawahiri, Omar Abdel-Rahman, and Abd al-Hamid Kishk. Zawahiri was released from prison in 1984, before traveling to Afghanistan and forging a close relationship with Osama Bin Laden.

Note the variety of features in action to make the different integrations here: *trial of … Islambouli* is integrated with *Islambouli and … were tried* mainly through lexical features, while *assassin Khalid Islambouli* is integrated with *Islambouli … carried out the assassination* through both lexical and dependency-based features. The multiple contributions from the different features and the presence of multiple high-likelihood integration points allow us to be confident that we have performed an appropriate integration.

Importantly, Figure 3 shows that integrating the new document not only reinforces existing beliefs, but also provides new information, e.g. the names of some of the people who were indicted in Islambouli's trial and the events they have participated in. New information not present in either text is also derived from the integration, in the form of new temporal relations. We follow the lead of (Mani et. al. 2006) in applying a temporal closure algorithm to our graph, based on a temporal relation transitivity table. This approach identifies a variety of temporal inferences. For example, since Zawahiri's indictment was *during* Islambouli's trial and Islambouli was executed *after* the trial, the algorithm can conclude that Zawahiri was indicted *before* Islambouli's execution. Finding and inferring such new information is the real point of performing graph integration; we process new documents not to be told what we already know, but to learn something new.

Conclusions

In this paper, we have presented an approach to machine reading that leverages a number of state-of-the-art natural language processing technologies to automatically populate a graph-based knowledge base. We have shown how to integrate a variety of shallow semantic parsing

techniques for basic units like entities, events, semantic roles and temporal relations into one cohesive semantic graph that better represents the meaning of the text. Entities and events form the nodes of these graphs, and entity coreference, semantic roles and temporal relations hold these nodes together.

We have also discussed our approach to integrating new documents with an existing semantic knowledge base: a machine learning model trained on features that identify similarities, both lexically and in the dependency structure, between the semantic structure of the new document and the semantic structure that has already been stored. By selecting multiple related entities and events for which these features predict high similarity, we can reach high levels of confidence in integrating these entities and events with those in the existing semantic structure.

Important areas of future research not addressed in this paper include the processing of causal, explanatory and some discourse relations, which are crucial for reasoning in applications like automated science tutors. Additionally, the integration of multiple documents can result in inconsistencies in the knowledge base, especially when the source of information is the web. This can be addressed by adding relational links to indicate contradictions and by decreasing the confidence estimates of associated relations. Conversely, corroborating information from multiple documents should increase confidence estimates.

Acknowledgments

This research was performed under an appointment of the first author to the Department of Homeland Security (DHS) Scholarship and Fellowship Program, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and DHS. ORISE is managed by Oak Ridge Associated Universities (ORAU) under DOE contract number DE-AC05-06OR23100. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of DHS, DOE, or ORAU/ORISE.

References

Bethard, S. and Martin, J. H. 2006. Identification of Event Mentions and their Semantic Class. In *Proceedings of Coling/ACL 2006*.

Grishman, R. and Sundheim, B. 1996. Message Understanding Conference-6: A Brief History. In *Proceedings of COLING 1996*, 466-471.

Hacioglu, K., Chen, Y. and Douglas, B. 2005. Automatic Time Expression Labeling for English and Chinese Text. In *Proceedings of CICLing-2005*.

Hacioglu, K., Douglas, B. and Chen, Y. 2005. Detection of Entity Mentions Occurring in English and Chinese Text. In *Proceedings of HLT-EMNLP 2005*.

Jiang, Z. P., and Ng, H. T. 2006. Semantic Role Labeling of NomBank: A Maximum Entropy Approach. In *Proceedings of EMNLP 2006*, 138-145.

Kintsch, W. 1998. *Comprehension: A Paradigm for Cognition*. Cambridge University Press, Cambridge.

Lin, D. and Pantel, P. 2001. Discovery of inference rules for Question Answering. In *Natural Language Engineering, 7*(4):343-360.

Mani, I., Verhagen, M., Wellner, B., Lee, C. M. and Pustejovsky, J. 2006. Machine Learning of Temporal Relations. In *Proceedings of Coling/ACL 2006*.

Nielsen, R., Ward, W. and Martin, J. H. 2006. Toward Dependency Path based Entailment. In *Proceedings of the Second PASCAL Recognizing Tetual Entailment Challenge Workshop*. Venice, Italy.

Nyberg, E., Mitamura, T., Frederking, R., Pedro, V., Bilotti, M., Schlaikjer, A. and Hannan, K. 2005. Extending the JAVELIN QA System with Domain Semantics. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005).*

Olsson, F. 2004. A survey of machine learning for reference resolution in textual discourse. SICS Technical Report T2004:02. Swedish Institute of Computer Science.

Pradhan, S., Hacioglu, K., Krugler, V., Ward, W., Martin, J. H. and Jurafsky, D. 2005. Support vector learning for semantic argument classification. *Machine Learning*, 60(1):11–39.

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Day, D., Ferro, L., Gaizauskas, R., Lazo, M., Setzer, A. and Sundheim, B. 2003. The TimeBank Corpus. *Corpus Linguistics*, 647-656.

Surdeanu, M., Harabagiu, S., Williams, J., and Aarseth, P. 2003. Using predicate-argument structures for information extraction. In *Proceedings of ACL 2003*.

Toutanova, K., Haghighi, A., and Manning, C. D. 2005. Joint learning improves semantic role labeling. In *Proceedings of ACL 2005*.