Knowledge-Based Labeling of Semantic Relationships in English

Alicia Tribble
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
atribble@cs.cmu.edu

Abstract

An increasing number of NLP tasks require semantic labels to be assigned, not only to entities that appear in textual elements, but to the relationships between those entities. Interest is growing in shallow semantic role labeling as well as in deep semantic distance metrics grounded in ontologies, as each of these contributes to better understanding and organization of text. In this work I apply knowledge-based techniques to identify and explore deep semantic relationships in several styles of English text: nominal compounds, full sentences in the domain of knowledge acquisition, and phrase-level labels for images in a collection. I also present work on a graphical tool for exploring the relationship between domain text and deep domain knowledge.

1 Introduction

As our command of NLP techniques has grown over the decades, the tasks which we can accomplish have become more useful and complex: we can (to an increasing extent) answer questions, create summaries, and even create new knowledge by extracting and merging facts from large text corpora. To make our systems reach their potential on these tasks, we need to extend our analysis of text into deep semantics, often grounded in world knowledge.

In this work, I explore the semantic relationships in several styles of English text using knowledge-driven NLP techniques as well as a novel graphical tool for the navigation of knowledge bases (KBs, or ontologies).

I begin by describing a system based on augmented LFG-style grammar rules, appropriate for the domain-limited sentences that are required for knowledge entry by knowledge base engineers. In a subsequent system for interpreting nominal compounds, I rely more heavily on the knowledge already stored in the knowledge base to guide a heuristic search for meaning (Tribble and Fahlman, 2006).

These systems demonstrate how a knowledge base can contribute to NLP performance. During development of the systems, knowledge acquisition and organization became important sub-topics. In response I began work on a graphical tool, SconeEdit (Tribble, Lambert, and Fahlman, 2006). SconeEdit allows users to navigate the semantic concepts and relations in a text corpus, guided by the rich, grounded features of these concepts in a knowledge base.

With this interface as a scaffold, future work entails improving the analysis systems for noun compounds and full sentences, and incorporating these systems in a comparative evaluation of the graphical and NLP-based methods for exploring semantic relationships in domain-restricted text. In addition, I will use this framework to evaluate a knowledge-

Proceedings of the NAACL-HLT 2007 Doctoral Consortium, pages 13–16,
Rochester, April 2007.

based approach for the task of retrieving labeled images from a collection.

2 Semantic Analysis for Knowledge Engineering

One of the motivating goals of this work is to leverage the power of NLP tools to ease the burden of knowledge engineers who develop ontological resources. By converting English sentences into a semantic representation automatically, a system provides an intuitive input method for adding new knowledge.

2.1 Knowledge Engineering in Scone

The context for this work is the Scone Knowledge Representation (KR) Project (Fahlman, 2005). The Scone KR System encompasses an inference engine along with a set of upper-level domain ontologies. As with other large KR systems along the lines of CYC (Lenat, 1986), knowledge engineers create much of the upper-level KB content by hand.

To develop a system that would address the needs of these engineers, I collected a corpus of English sentences covering the six core structure-building tasks in Scone:

- Defining a type
- Adding an instance of a type
- Defining a relation between types
- Adding an instance of a relation
- Defining a new role (HAS-A) relation
- Instantiating a role-filling relation

2.2 A Grammar-Based System

The resulting corpus displayed a high degree of semantic cohesion, as expected, but with a wide degree of syntactic variation. To transform these sentences automatically into the Scone KR, I developed a set of semantic interpretation functions and added them as callouts in an existing LFG-style syntactic grammar. The resulting augmented English grammar is applied to new sentences using the LCFlex parser of Rosé and Lavie (2000). In this way, every parse constituent can be conditioned on queries to the knowledge base, allowing not only flat semantic features (e.g. "is the noun animate?") but rich structural knowledge ("does this person own a pet?") to be applied during the parse.

The new grammar rules produce output in the Scone KR formalism. As a result, the output can be read as the knowledge-grounded meaning of an input sentence, and it can also become additional input to the Scone inference engine, adding to the store of background knowledge or making a new query. However, the appeal of this design is limited by the fact that, as in many grammar-based systems, the rules themselves are costly to write and maintain.

2.3 Adding Generalization

For this reason, I modified the approach and examined the effectiveness of a few general "preference" rules, based on syntax. In contrast with the grammar system, the search for interpretations can now be driven, rather than pruned, by domain knowledge. I tested this approach on the interpretation of noun compounds, where the lack of syntactic cues requires heavy reliance on semantic interpretation (Tribble and Fahlman, 2006). I found that a majority of compounds, even in a new textual domain, could be analyzed correctly using the new set of rules along with an appropriate domain-specific KB.

3 A Graphical Tool for Exploring Semantic Relationships

While the cost of grammar writing can be reduced with updated algorithms, developing and maintaining large knowledge repositories is one of the key challenges in knowledge-based NLP: the knowledge acquisition "bottleneck". My hypothesis is that a natural-language (NL) interface is an important tool for easily modifying and adding knowledge in a complex KR system like Scone; language is an intuitive way for users to express what they want from the knowledge base.

In the course of developing NL tools for the Scone Project, I also recognized the need to view domain text, domain knowledge, and the semantic relationships that they share in a "snapshot". Integrating textual and graphical exploration gives users a comfortable handle on the knowledge base, even when they don't know exactly what they want.

I designed the SconeEdit knowledge- and text-browsing tool (Tribble, et al. 2006) in response to this need. The tool provides an annotated view of text chosen by the user, allowing him to see what concepts and vocabulary from the text are currently in the KB. Alongside this Text View, SconeEdit provides a navigable snapshot of the knowledge base (KB View), centered on concepts that appear in the text. This unified browser establishes a principled, coverage-driven way to "surf" the KB and add new knowledge. A screenshot of SconeEdit, recently updated to view images as well as text, is shown in Figure 1.

The SconeEdit tool has already been used by groups outside the Scone Project, for the purpose of qualitatively evaluating knowledge bases for use in new subdomains. My goal for the conclusion of this work is to synergize the lines of research described so far, building our English analysis tools into the SconeEdit interface. With the resulting tool I can run a detailed evaluation of my English analyzers, as well as shed light on the usability of text-based versus graphical knowledge entry.

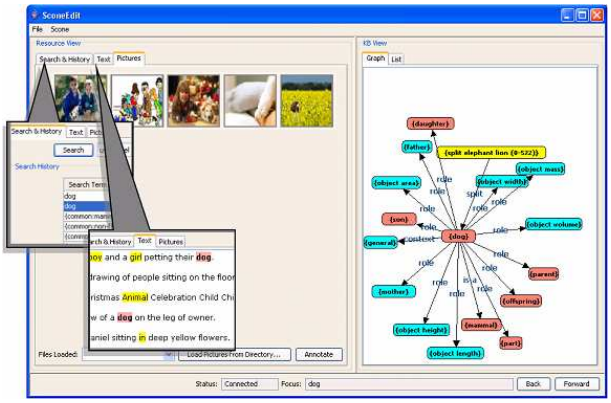

Figure 1. Screenshot of SconeEdit, updated to display images as well as text.

4 Task-Based Evaluation: Retrieving Labeled Images

To bring this work to bear in a task-based evaluation, I have also started developing a system for labeled image retrieval. To retrieve images of interest from large collections, traditional systems rely on matching between a high-level query and low-level image features, or on matching the query with an unordered bag-of-words that has been attached to each image. In current work I am investigating sentence fragments, which retain some syntactic structure, as a useful style of image annotation that is complementary to the current bag-of-words style. Analysis of 2,776 image titles downloaded from the web establishes that fragment-style labels are intuitive, discriminative, and useful.

These labels can be used to retrieve images from a collection in the following way: first, a typed query is given to the system (e.g. "people petting their dogs"). An English analyzer, using improvements to the techniques described in Section 2, produces the Scone semantic representation of this query (a semantic graph). Next, the Scone inference engine is used to match the query against pre-computed semantic representations of the image labels. The system retrieves the image whose label matches best. Figure 2 is an example retrieved for this query by Google Image Search.


Figure 2. Image retrieved by Google Image Search for "people petting their dogs".

4.1 Development Data

In order to train the functions that measure a "match" in the knowledge base, as well as to improve the English-to-Scone analysis, I need training data in the form of images, their fragment-style labels, and one or more query that matches each image and its label.

I collected one corpus of images with their fragment-style labels from the publicly available collection on Flickr (http://www.flickr.com). A second corpus of fragment-labeled images has been provided by one the authors of von Ahn and Dabbish (2004). In many cases, a single image has

multiple fragment-style labels. To convert this data into the format I need, I can use the redundant labels as substitute "queries", under the assumption (which should be validated experimentally) that image-retrieval queries often take the form of sentence fragments, as well.

An evaluation that uses these labels for image retrieval will proceed as follows: A subset of the labeled images which were not seen or used in previous work will be reserved as test data. Remaining images with their labels and queries will be used to improve the English-to-Scone analysis system and the semantic similarity functions within Scone. Finally, the queries for the test set will be submitted to the retrieval system, and system results will be compared to the "correct" images given by the test set. Precision and recall can be calculated under a variety of conditions, including one-image-per-query and several-images-per-query. Comparison to shallow techniques for label matching, as used with bag-of-words style labels, will also be a feature of this evaluation.

5 Conclusion

In summary, I have presented a body of work on exploring and labeling the deep semantic relationships in English text. A grammar-based system for sentences and a heuristic search system for noun compounds explore the role of domain knowledge in tools for syntactic and deep semantic analysis. In addition, I designed and demonstrated graphical tool for exploring rich semantic features in text, grounded in a knowledge base or ontology. The tool has been used by our own knowledge engineers as well by other research teams at CMU.

I will build on this work in the coming months as I prepare for two evaluations: a study on the usability of natural language and graphical tools for navigating a knowledge base, and a task-based evaluation on labeled image retrieval. These evaluations should bring closure to the work as a contribution in the field of semantic analysis of text.

References

Scott E. Fahlman. 2005. The Scone User's Manual. *http://www.cs.cmu.edu/~sef/scone.*

Alicia Tribble, Benjamin Lambert and Scott E. Fahlman. 2006. SconeEdit: A Text-Guided Domain Knowledge Editor. In *Demonstrations of HLT-NAACL 2006*. New York.

Alicia Tribble and Scott E. Fahlman. 2006. Resolving Noun Compounds with Multi-Use Domain Knowledge. In *Proceedings of FLAIRS-2006*. Melbourne Beach, Florida.

Alicia Tribble and Carolyn P. Rosé. 2006. Usable Browsers for Knowledge Acquisition. In *Proceedings of CHI-2006*. Montreal, Quebec.

Carolyn P. Rosé and Alon Lavie. 2001. Balancing Robustness and Efficiency in Unification-Augmented Context-Free Parsers for Large Practical Applications. In J.C. Junqua and G. Van Noord, eds. *Robustness in Language and Speech Technology.* Kluwer Academic Press.

D. B. Lenat, M. Prakash and M. Shepherd. 1986. Cyc: using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks.. In *AI Magazine.* 6:4.

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of ACM CHI* (pp 319—326).