Supervised and Unsupervised Models for Propositional Analysis

Simon Dennis[*1]
Dan Jurafsky[*#]
Dan Cer[#]

Institute of Cognitive Science[*]
Centre for Spoken Language Research[#]
University of Colorado

Introduction

At the level of lexical semantics, distributional methods that generate word representations from large text corpora have proven to be useful practical techniques (Landauer, 2002) as well provocative theoretical models (Landauer & Dumais, 1997). Following in the footsteps of Latent Semantic Analysis (LSA, Deerwester, Dumais, Furnas, Landauer, & Harshman, 1991) there has been substantial interest in these techniques, and we are currently aware of some eighteen published methods for deriving semantic content (including Blei, Ng, & Jordan, 2002; Dennis, 2003; Griffiths & Steyvers, 2002; Hofmann, 2001; Lin & Pantel, 2001; Redington, Chater, & Finch, 1998)[2]. Similarly, there has also been interest in unsupervised methods for syntactic analysis (Klein & Manning, 2001; Magerman & Marcus, 1990) and the performance of current methods as measured by the recall and precision of constituent identification has been gradually improving.

It has long been realized, however, that many tasks require representations at a propositional level beyond that captured by lexical semantics and purely syntactic analysis (Kintsch, 1998). A **proposition** refers to a unit consisting of a relational term (the predicate) and one or more arguments with specific semantic roles. Propositions have been important in the work of both linguists (Beirwisch, 1969; Fillmore, 1968; van Dijk, 1972) and psychologists (Kintsch, 1974).

Despite more than twenty years of research, however, the problem of how to automatically map text into propositional structures has not been solved (Dennis & Kintsch, submitted). Researchers in the area of text comprehension have devised sets of guidelines that allow text to be hand coded in a fairly consistent way, but this is not an adequate solution and places restrictions on the size and range of texts that can be utilized.

In addition, the ability to extract propositions from text is crucial in a number of applied domains including question answering and information extraction. The majority of current information extraction systems are based on a surface analysis of text applied to very large textbases. Whereas the dominant approaches in the late 1980s and early 1990s would attempt linguistic analysis, proposition extraction and reasoning, most current systems look for answer patterns within the raw text and apply simple heuristics to extract relevant information (Brill, Dumais, & Banko, ; Brill, Lin, Banko, Dumais, & Ng, ; EAGLES, 1998; Voorhees, 2002). Such approaches have been shown to work well when information is represented redundantly in the textbase and when the type of the answer is unambiguously specified by the question and tends to be unique within a given sentence or sentence fragment. While these conditions often hold for general knowledge questions of the kind found in the Text Retrieval Conference (TREC) Question Answer track, there are many intelligence applications for which they cannot be guaranteed. Often relevant information will only be stated once or may only be inferred and never stated explicitly. Furthermore, the results of the most recent TREC Question Answer competition suggest that reasoning systems may now have reached a level of sophistication that allows them to surpass

[1] Funding for this work was provided by NSF grant IIS-0325646, NSF grant EIA-0121201 and the US Department of Education grant R305G020027.
[2] For a list of papers describing many of these techniques see http://lsa.colorado.edu/~simon/LexicalSemantics/

the performance possible using surface based approaches. In the 2002 TREC competition, the PowerAnswer system (Moldovan et al., 2002), which converts both questions and answers into propositional form and employs an inference engine, achieved a confidence weighted score of 0.856, a substantive improvement over the second placed exactanswer (Soubbotin & Soubbotin, 2002), which received a score of 0.691 in the main question answering task. The propositions employed by the PowerAnswer system, however, were restricted to relatively simple relationships and it may be that with more sophisticated propositional analysis performance could be significantly improved.

For both theoretical and practical reasons, then, the search for a robust and comprehensive method for proposition extraction is crucial. In this paper, we outline a series of attempts to achieve this goal including supervised methods for identifying and labeling semantic roles, an unsupervised method that performs reasonably well in some domains and, finally, preliminary work aimed at broadening the range of domains to which our unsupervised method applies.

Supervised Propositional Analysis

The task addressed by supervised semantic parsers is to take a sentence and assign role labels (Agent, Patient, Instrument, Location, etc) to the relevant constituents for each of the predicates in the sentence (Blaheta & Charniak, 2000; Gildea & Jurafsky, 2002; O'Hara & Wiebe, 2002; Palmer, Rosenzweig, & Cotton, 2001).

For instance, given the sentence:

Sampras outguns Agassi in Utah

a system would produce an annotation such as:

[$_{\text{Agent}}$ Sampras] outguns [$_{\text{Patient}}$ Agassi] [$_{\text{Location}}$ in Utah]

This work has been driven, at least in part, by the availability of semantically labeled corpora such as Propbank (Kingsbury, Palmer, & Marcus, 2002) and FrameNet (Fillmore, Wooters, & Baker, 2001) which provide the relevant training data.

In our system (Gildea & Jurafsky, 2002; Pradhan, Hacioglu, Ward, Martin, & Jurafsky, 2003), for each training sentence, we first syntactically parse the sentence, extract various syntactic, lexical, and semantic features, and train a classifier. The classifier looks at each word chunk or parse-constituent in the sentence, and decides for each one whether it is an argument of the verb, and if so what its semantic role is. These features include the syntactic type of the constituent (NP, PP, S, VP, etc), the identity of the verb, the path in the parse tree between the verb and the constituent, whether the constituent is before or after the verb, etc. We have used various classifier architectures, including support vector machines and decision trees. Given enough training data, our classifiers achieve quite reasonable accuracy on semantic role labeling, although there is definitely still room for improvement.

One important limitation of supervised semantic parsers is that they rely on the accuracy, coverage and labeling scheme of their training set. There are a great many schemes that have been proposed ranging in granularity from very broad, such as the two macro-role proposal of Van Valin (1993), through theories that propose nine or ten roles, such as Fillmore (1971), to much more specific schemes that contain domain specific slots such as ORIG_CITY, DEST_CITY or DEPART_TIME that are used in practical dialogue understanding systems (Stallard, 2000). It is well-known to be difficult to label semantic roles in a general way, and each database (like FrameNet and PropBank) makes choices about roles that are often incompatible. Furthermore, our system's performance is quite dependent on the training data looking like the test data. Therefore we think it is key to understand how this task could be extended to an unsupervised one.

Unsupervised Semantic Parsing

The task addressed by unsupervised semantic parsers is to create proposition-like units without recourse to labeled training data. Furthermore, we would like to minimize our reliance on syntactic features derived from knowledge rich sources such as treebanks or grammars. The reasons for interest in this task are two fold. Firstly, at a theoretical level the information that can be induced by our system forms a lower bound on what a human inductive system might achieve. People do not have access to semantically or syntactically labeled training data when learning a language and so in order for our systems to be relevant to the debate on what can and cannot be induced it is crucial that our system be subject to the same constraints. Secondly, as indicated above, supervised semantic parsers tend to degrade significantly when applied to text that differs from that on which they were trained. Creating sufficiently large corpora of labeled training data is expensive, error prone and time consuming, however.

The key to our attempts at unsupervised parsing is a switch from intentional to extensional representation. In systems that employ intentional semantics, as above, the meanings of representations are defined by their intended use and have no inherent substructure. The names of the roles are completely arbitrary and carry representational content only by virtue of the inference system in which they are embedded.

Now contrast the above situation with an alternative *extensional* representation of “Sampras outguns Agassi”, in which roles are defined by enumerating exemplars, as follows:

Sampras: Kuerten, Hewitt
Agassi: Roddick, Costa

The winner role is represented by the distributed pattern of Kuerten and Hewitt, words that have been chosen because they are the names of people who have filled the “X” slot in a sentence like “X outguns Y” within the experience of the system. Similarly, Roddick and Costa are the names of people that have filled the “Y” slot in such a sentence and form a distributed representation of the loser role.

The use of extensional semantics, of this kind, has a number of advantages. First, defining a mapping from raw sentences to extensional meaning representations is much easier than defining a mapping to intentional representations because it is now only necessary to align sentence exemplars from a corpus with the target sentence. The difficult task of either defining or inducing semantic roles is avoided.

Second, because the role is now represented by a distributed pattern it is possible for a single role vector to simultaneously represent roles at different levels of granularity. The pattern {Kuerten, Hewitt} could be thought of as a proto-agent, an agent, a winner, and a winner of a tennis match simultaneously. The role vectors can be determined from a corpus during processing, and no commitment to an a priori level of role description is necessary.

Third, extensional representations carry content by virtue of the other locations in the experience of the system where those symbols have occurred. That is, the systematic history of the comprehender grounds the representation. For instance, we might expect systematic overlap between the winner role and person-who-is-wealthy role because some subset of {Kuerten, Hewitt} may also have occurred in an utterance such as “X is wealthy”. These contingencies occur as a natural consequence of the causality being described by the corpus and have been dubbed *inference by coincidence* (Dennis, in press-b).

To create extensional representations of sentences, Dennis (in press-a; in press-b) used String Edit Theory (Sankoff & Kruskal, 1983) to align sentences from a corpus with the target sentence.

As an illustrative example, suppose that we wish to create a representation for the sentence “Sampras outguns Agassi in Utah” and that the following are the most probably alignments across a corpus:

	Sampras			**outguns**		**Agassi in Utah**
Gustavo	Kuerten			downs	Andy	Roddick
	Hewitt	advances to the finals		defeating		Costa

The extensional representation thus formed would be:

Sampras: Kuerten, Hewitt
Outguns: downs, defeating
Agassi: Roddick, Costa

To test the model, sixty nine articles were taken from the Association of Tennis Professionals (ATP) website at http://www.atptennis.com/. The articles were written between September 2002 and December 2002 and ranged in length from 134 to 701 words. In total there were 21212 words in the corpus. The documents were manually divided into sentences and the mean sentence length was 23.7.

Then 377 questions of the form "Who won the match between X and Y? X" were created. Any result that could be deduced from the article text was included. So, for instance, results that required the resolution of an anaphoric reference from other sentences in the same document were retained. Also, the winner was alternated between the first and second name positions so that the model could not simply repeat the name in the first slot in order to answer the question.

Finally, the model was presented with the same sorts of questions with the answer omitted. An extensional representation of each question was created and matched against the extensional representations of the sentences from the target articles to identify sentences that might contain the critical information. From these sentences the filler most likely to be appropriate in the answer slot was chosen (see Dennis, in press-b for details).

On 67% of occasions the model correctly returned the winner of the match. 26% of the time it incorrectly produced the loser of the match. 5% of the time it responded with a player other than either the winner or loser of the match and on 3% of occasions it committed a type error, responding with a word or punctuation symbol that was not a player’s name. Interestingly, when the model was correct, on 41% of occasions the sentence that it chose as the most likely to contain relevant information was one from which an answer could be inferred but that did not contain a literal statement of the result (see Table 1 for examples). On 30% of occasions the most probably sentence contained a literal statement of the result, and on 29% of occasions it selected a sentence from which the answer did not directly follow. That is, inference by coincidence plays a significant role in the performance of the model.

While there is clearly room for improvement in the performance of the model, it demonstrates that it is possible to extract proposition-like representations from open text using only simple string edit operations that have no built in grammatical or semantic knowledge. Furthermore, the importance of inference by coincidence in the model suggests that systems based on intentional representational systems may be throwing away a critical source of statistical information that may underpin the robustness of the human comprehension apparatus.

Table1: Examples of inference by coincidence

Who won the match between Carlsen and Kiefer ? Carlsen

Kafelnikov now meets Kenneth Carlsen of Denmark in the second round.

Who won the match between Kiefer and Safin ? Safin

Safin , Kafelnikov surge toward hometown showdown

Who won the match between Voltchkov and Haas ? Voltchkov

According to Haas, the injury first arose during Wednesday's match against Sargis Sargsian, and became progressively worse during practice and then the match against Voltchkov.

Unsupervised Domain-General Semantic Parsing

In the work presented in the previous section, the semantic roles that were required to be extracted were quite specific – namely we needed to be able to distinguish between winners and losers. Furthermore, fillers of the roles that the model had to extract were always single words. As SET has no notion of constituent, the question arises as to how well it will perform when more general semantic roles are required that take multiword fillers.

To test the model, we took sentences from Propbank that contained 10 or less words once their labeled constituents had been reduced to their heads and induced role vectors for the constituents labeled agent, actor1, patient and theme as above. Partitioning around medoids (PAM, Kaufman & Rousseeuw, 1990) was then used to create two clusters of the resulting role vectors under the assumption that the constituents labeled agent and actor1 should form a protoagent cluster, while the constituents labeled patient and theme should form a protopatient cluster. Identifying the cluster with the highest proportion of protoagents as the protoagent cluster (and the other cluster as the protopatient cluster), the role vectors were correctly labeled on 75% of occasions. While this result is not unreasonable given that the algorithm is not being provided with any linguistic knowledge, a baseline where constituents that occur before the verb are classified as protoagents and constituents that occur after the verb are classified as protopatients performs as well (76%). Furthermore, this baseline can be increased to 84% if we also conditionalize on the identity of the verb.

A close examination of the tennis domain suggests that the performance on that task may have been facilitated by the fact that role information can often be deduced from the words that immediately surround a given target word. In general, however, this is not guaranteed to be the case especially for higher level semantic roles.

In current work we are investigating whether the performance can be improved by incorporating more linguistic knowledge. In particular, we are investigating how well the system could perform if it has access to knowledge of the dependencies between predicates and arguments as provided by a dependency parser. In this way, we are separating the task of inducing simple syntactic dependency links from the task of semantic role induction.

Initial results were obtained by first parsing the entire Propbank training set using the Minipar dependency parser (Lin, 1993, 1994, 1998). From these parses, we constructed role vectors by collecting all of the words that appeared at the terminal of each distinct path from a target verb to a labeled constituent (probability weighted). Further, the role vectors were conditioned on the presence of paths from the target verb to other labeled constituents.

Consider the following sentences:

[$_{protopatient}$ The butter] melted.
[$_{protoagent}$ The chef] melted [$_{protopatient}$ the butter].

Even though 'butter' in the first sentence has the same path to the target verb as 'chef' in the second, these words would contribute to different role vectors as the path from the verb to the protopatient appears only in the second sentence.

By clustering role vectors constructed in this manner, we are able to obtain a classification accuracy of 75%. However, for the dataset used to generate the clusters, the baseline based upon whether a labeled constituent occurs before or after the target verb is 79%.

Preliminary error analysis indicated that the system was not performing well on sparse role vectors. To assess to what extent sparsity compromises performance, we re-clustered the role vectors using only those that had 25 or more entries. This resulted in classification accuracy of 93%. The positional baseline for the data set used in this latter clustering experiment is 90%. This result suggests that classification performance on the larger data set may be increased by smoothing the representations for the sparser role vectors. Further, we suspect that enriching the representation of the items being cluster by including path and target verb based features will improve system performance.

Conclusions

In this paper, we have briefly outlined three systems for extracting propositions from text – one supervised and two unsupervised. In order do unsupervised role induction it was necessary to shift from the intentional representational system typically employed in linguistic analysis and supervised semantic parsers to an extensional system. We believe that this move will not only facilitates the mapping from text to propositions, but will also lead to more robust inferencing systems that take advantage of inference by coincidence.

Performance in our unsupervised systems still lags that of supervised parsers, however, especially when more general semantic roles (like protoagent and protopatient) must be distinguished. Initial results using a dependency parser suggest that performance can be improved using more linguistically informed features. Furthermore, we believe that it will be possible to induce dependency relationships to create a completely unsupervised mechanism.

References

Beirwisch, M. (1969). On certain problems of semantic representation. *Foundations of language, 5*, 153-184.

Blaheta, D., & Charniak, E. (2000). *Assigning function tags to parsed text.* Paper presented at the Proceedings of the 1st Annual Meeting the North American Chapter of the ACL (NAACL), Seattle, Washington.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). *Latent Dirichlet Allocation.* Paper presented at the Nueral Information Porcessing Systems.

Brill, E., Dumais, S., & Banko, M.*An analysis of the AskMSR question-answering system.*, from http://research.microsoft.com/~sdumais/EMNLP_Final.pdf

Brill, E., Lin, J., Banko, M., Dumais, S., & Ng, A.*Data-intensive question answering*, from http://www.at.mit.edu/people/jimmylin/publications/Brill-etal-TREC2001.pdf

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1991). Indexing By Latent Semantic Analysis. *Journal of the American Society For Information Science, 41*, 391-407.

Dennis, S. (2003). *A comparison of statistical models for the extraction of lexical information from text corpora.* Paper presented at the Twenty Fifth Conference of the Cognitive Science Society.

Dennis, S. (in press-a). A memory-based theory of verbal cognition. *Cognitive Science*.
Dennis, S. (in press-b). An unsupervised method for the extraction of propositional information from text. *Proceedings of the National Academy of Sciences*.
Dennis, S., & Kintsch, W. (submitted). The text mapping and inference generation problems in text comprehension. In.
EAGLES. (1998). *Preliminary recommendations on semantic encoding: Information Extraction*, from http://www.ilc.pi.cnr.it/EAGLES96/rep2/node30.html
Fillmore, C. J. (1968). The case for case. In E. Black & R. T. Harms (Eds.), *Universals of linguistic theory*. New York: Holt, Reinhardt and Winston.
Fillmore, C. J. (1971). Some problems for case grammar. In R. J. O'Brien (Ed.), *22nd round table. linguistics: deelopments of the sixties - viewpoints of the seventies* (Vol. 24, pp. 35-56). Washington D. C.: Georgetown University Press.
Fillmore, C. J., Wooters, C., & Baker, C. F. (2001). *Building a large lexical databank which provides deep semantics.* Paper presented at the Proceedings of the Pacific Asian Conference on Language, Information and Computation, Hong Kong.
Gildea, D., & Jurafsky, D. (2002). Automatic Labeling of semantic roles. *Computational Linguistics, 28*(3), 245-288.
Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In C. D. Schunn & W. D. Gray (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society*: LEA.
Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42*(1-2), 177-196.
Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.
Kingsbury, P., Palmer, M., & Marcus, M. (2002). *Adding semantic annotation to the Penn TreeBank.* Paper presented at the Proceedings of the Human Language Technology Conference, San Diego, CA.
Kintsch, W. (1974). *The representation of meaning in memory*. New York: Wiley.
Kintsch, W. (1998). *Comprehension: a paradigm for cognition*: Cambridge University Press.
Klein, D., & Manning, C. D. (2001). *Distributional phrase structure induction.* Paper presented at the The fifth conference on natural language learning.
Landauer, T. K. (2002). *Applications of Latent Semantic Analysis.* Paper presented at the 24th Annual Conference of the Cognitive Science Society, Fairfax, VA.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.
Lin, D. (1993). *Princinciple-based parsing without overgeneralization.* Paper presented at the Proceedings of ACL-93, Columbus, OH.
Lin, D. (1994). *Principar - an efficient, broad-coverage, principle-based parser.* Paper presented at the Proceedings of COLING-94, Kyoto, Japan.
Lin, D. (1998). *Dependency-based evaluation of MINIPAR.* Paper presented at the Workshop on the Evaluation of Parsing Systems, Granada, Spain.
Lin, D., & Pantel, P. (2001). *Induction of semantic classes from natural language text.* Paper presented at the ACM SIGKDD Conference on Knowledge Discovery and Data Mining.
Magerman, D. M., & Marcus, M. P. (1990). *Parsing a natural language using mutual information statistics.* Paper presented at the National Conference on Artificial Intelligence.
Moldovan, D., Harabagiu, S., Girju, R., Morarescu, P., Lacatusu, F., Novischi, A., et al. (2002). *LCC tools for question answering.* Paper presented at the Eleventh Text Retrieval Conference (TREC 2002).
O'Hara, T., & Wiebe, J. (2002). *Classifying preposition semantic roles using class-based lexical associations* (No. NMSU-CS-2002-013): Computer Science Department, New Mexico State University.
Palmer, M., Rosenzweig, J., & Cotton, S. (2001). *Automatic predicate arguemnt analysis of the Penn TreeBank.* Paper presented at the Proceedings of HLT 2001, First International Conference on Human Language Technology Research, San Francisco.

Pradhan, S., Hacioglu, K., Ward, W., Martin, J., & Jurafsky, D. (2003). *Shallow semantic parsing using support vector machines* (No. TR-CSLR-2003-03): Center for Spoken Language Research, University of Colorado.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*(4), 425-469.

Sankoff, D., & Kruskal, J. B. (1983). *Time warps, string edits and macromolecules: the theory and practise of sequence comparison*: Addison Wesley.

Soubbotin, M. M., & Soubbotin, S. M. (2002). *Use of patterns for detection of answer strings: A systematic approach.* Paper presented at the Eleventh Text Retrieval Conference (TREC 2002).

Stallard, D. (2000). *Talk'n'travel: A conversational system for air travel planning.* Paper presented at the Proceedings of the 6th Applied Natural Language Processing Conference (ANLP'00).

van Dijk, T. A. (1972). *Some aspects of text grammars*. The Hague: Mouton.

Van Valin, R. D. (1993). A synopsis of role and reference grammar. In R. D. Van Valin (Ed.), *Advances in Role and Reference Grammar*. Amsterdam: John Benjamins Publishing Company.

Voorhees, E. M. (2002). *Overview of the TREC 2002 Question Answering Track.* Paper presented at the Eleventh Text Retrieval Conference (TREC 2002).