Adding syntactic information to LSA

Peter Wiemer-Hastings
Peter.Wiemer-Hastings@ed.ac.uk
School of Cognitive Science / ICCS, Division of Informatics
University of Edinburgh, Edinburgh EH8 9LW Scotland

Abstract

Much effort has been expended in the field of Natural Language Understanding in developing methods for deriving the syntactic structure of a text. It is still unclear, however, to what extent syntactic information actually matters for the representation of meaning. LSA (Latent Semantic Analysis) allows you to derive information about the meaning without paying attention even to the order of words within a sentence. This is consistent with the view that syntax plays a subordinate role for semantic processing of text. But LSA does not perform as well as humans do in discriminating meanings. Can syntax be the missing link that will help LSA? This paper seeks to address that question.

Introduction

In the beginning, there was syntax. And it was good. But it did not give us what we really want to know about a text — what it means. Then there was latent semantic analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990, LSA), which provided a means of comparing the "semantic" similarity between a source and target text, and thereby giving some idea of meaning of the source. That was good too, almost as good as humans in a simple task, but not quite. Because LSA pays no attention to syntax at all — not even word order — one promising approach to improving LSA is by giving it some of the information that is provided by syntax. Knowledge about the syntactic structure of a sentence provides information about the relationships between the words: which words modify which other words, and the relationships between verbs and their arguments or thematic roles. The research presented here is an attempt to evaluate the benefits of providing LSA with thematic role information which comes from syntactic knowledge.

Previous work

The primary goal of the AutoTutor project (Graesser, Franklin, Wiemer-Hastings, & the Tutoring Research Group, 1998; Wiemer-Hastings, Wiemer-Hastings, & Graesser, 1999) is to model human tutorial dialogue. It is based on studies of the discourse patterns of human tutors during tutoring sessions (Person, Graesser, Magliano, & Kreuz, 1994). These analyses have shown that human tutors do not have complete understanding of their students' answers to questions, but the do get an approximation. For AutoTutor, LSA provides such approximate understanding of student inputs by comparing them to expected answers, and using the LSA cosines as a metric of the extent to which the student entered what was expected.

We evaluated this approach by randomly selecting a set of 8 student answers to each of 24 questions in our domain of computer literacy (Wiemer-Hastings, Graesser, Harter, & the Tutoring Research Group, 1998). We asked human raters to evaluate these answers by providing an aggregate measure of the percentage of student answer propositions that "match" some expected answer proposition. Proposition was defined loosely as an atomic sentence. Match was left to the human raters to define. Then we performed the same analysis with LSA, modeling the match function by adjusting the cosine threshold. The best performance was realized with a 200-dimensional space with a cosine threshold of 0.5. This provided a correlation of $r = 0.49$ with the average rating of the human judges. Because the distribution of ratings was skewed, we also calculated Cronbach's alpha. The average alpha score between human raters was $\alpha = 0.76$. The alpha score between LSA and the average human rating was $\alpha = 0.60$. These results were very encouraging. LSA provided much of the discrimination shown by human raters, enough to use in the AutoTutor system. It could however, be improved.

The obvious information source that LSA ignores is syntax. It is a "bag-of-words" approach, simply adding together term vectors to make a vector for a text. This paper is an attempt to identify whether the addition of syntactic knowledge can strengthen LSA judgments.

Related work

Partially as a result of the Behaviorist movement in psychology, linguistics and natural language processing focused for a long time primarily on the syntactic structure of sentences (Chomsky, 1981, for example). In the 70's and 80's, Schank sought to change this by claiming that semantics alone was sufficient (Schank & Riesbeck, 1981, for example). More recently, researchers from psychology have championed LSA as both a technique for determining the meaning of texts and as a model of human language.

Much of the recent interest (and controversy) regarding LSA can be traced to Landauer, Kintsch, and colleagues. They imported LSA from the realm of information retrieval and hailed it as part or parcel of a psy-

chological model of language understanding. Landauer and Dumais (1997) described LSA as a model of human language acquisition, using it to explain how the pace of lexical acquisition apparently outstrips the exposure to new words. Landauer has gone on to claim that LSA is a complete model of language understanding (Landauer, Laham, Rehder, & Schreiner, 1997). He explains away the existence of syntax by suggesting that it is only there to simplify the computational complexity of getting the words into an LSA-like representation in the first place.

Other psychologists have stressed the role which syntax can play in lexical acquisition. The syntactic bootstrapping(Gleitman & Gillette, 1994) theory shows how pre-verbal children can use their knowledge of syntax to help guide their acquisition of verbs.

Kintsch (1998) has appended LSA to his Construction/Integration model of text understanding as the semantic component. LSA provides a sort of spreading activation-like inclusion of related concepts when new information is integrated into a knowledge structure. This allows the system to perform a type of inference, making, for example, "driver" and "computer" available when "bus" is mentioned in a text.

In other related psychological approaches, MacDonald has proposed a used a variant of LSA to predict semantic priming (McDonald, 2000). And Ramscar and colleagues have used LSA to model analogical reasoning (Packiam-Alloway, Ramscar, & Corley, 1999).

The HAL system (Burgess & Lund, 1997) is similar to LSA in the sense that it is based on co-occurrences, but word order information enters the representation space through a weighting mechanism: A co-occurrence is weighted more heavily the fewer words intervened between the two words, within a window of usually ten words. So, two words that co-occur in immediate adjacency are weighted most strongly. This is not syntax, but it does grant some sensitivity to word order.

Burgess and Lund replicated earlier work by Finch and Chater (Finch & Chater, 1992) which showed that by applying a high-dimensional method to clustering the co-occurences of words in a corpus, it is possible to infer lexical categories that correspond well with standard syntactic theories. Finch and Chater also showed that you could use these categories to infer basic grammatical rules (see also (Siskind, 1996; Christiansen & Chater, 1999) for other corpus-based approaches to acquiring such information). Thus, there seems to be sufficient information in a corpus of text to statistically infer something about the syntactic structure of that corpus.

This does not mean, however, that a technique like LSA already has the type of syntactic information that we are attempting to incorporate here. For any particular sentence, LSA creates a vector just based on the bag of words that are in that sentence. It has no information about the word order within that sentence or about the relationships between the words.

Approach

Our initial success with LSA and the potential for improvement led us to examine how additional information could be provided. One obvious possibility is to use more classical natural language understanding techniques as a pre-filter for LSA. The idea is to use parsing, anaphora resolution and other dialogue-processing techniques to prepare chunks of text for LSA to process semantically. Alternatively, this could be viewed as using LSA as the semantic component of a classical natural language understanding system.

We preprocessed the student sentences and the expected answer sentences in the following way: First, we performed a basic syntactic segmentation of the sentences. Although there are surface-level parsing methods generally available (Abney, 1996, for example), their grammars must be modified to conform to the application. If this approach is successful, we will move to automated methods. For this test, we simply separated the sentences into atomic clauses or propositions, and then segmented them by hand, breaking them down into strings which corresponded to:

- subject noun phrase
- verb, including adverbs and adverbial phrases
- object noun phrase (when applicable)

This provides two types of additional information:

1. the grouping of words which belong together into "components"
2. the pseudo-semantic role of the components as derived from syntactic argument structure

Second, we resolved anaphora in the sentences, replacing pronouns by their antecedents. Finally, when there was a conjunction, we distributed the arguments. For example, if there was a sentence like, "Subject verb object1 and object2", it was broken into ("verb" "Subject" "object1") and ("verb" "Subject" "object2"), using a verb-prefix notation.

We made no attempt to do any other processing based on discourse relations for two reasons. First, LSA normally ignores "stop words" like "if" and "because" anyway. Second, extracting any more complex discourse relations would require the use of semantic understanding which is the goal of this process. Table 1 gives some some examples of sentences and their representations in this scheme.

There are three competing hypotheses of the effect on similarity judgments of using this additional information along with LSA:

1. Component grouping will increase discrimination because it adds information — the role of different components.
2. Component grouping will hurt discrimination because LSA works better on longer strings.
3. Component grouping will hurt grouping due to some complexity of combining individual component similarity scores.

Table 1: Example sentences and their representations

RAM stores the instructions to your programs.	("stores" "RAM" "the instructions to your programs")
If the new motherboard uses the same type of RAM, you can just take the SIMMs out of your old motherboard and install them in your new motherboard.	("if uses" "the new motherboard" "the same type of RAM") ("can just take out of your old motherboard" "you" "the SIMMs") ("and install in your new motherboard." "you" "the SIMMs")

The following section describes our first attempt to test these hypotheses using a straightforward combination of the between-component cosines.

Experiment 1

Given this type of representation, there remain a variety of ways to calculate the overall similarity between propositions based on the similarities of the components. In experiment one, we took the most straightforward approach, simply averaging the cosines of the respective components. In other words, we calculated the LSA cosine between the verb string from a student proposition and the verb string from an expected answer. We repeated this for the other sentence components. If there was an object string for one sentence and not for the other, a component score of zero was recorded. Then we averaged across the (normally two or three) components of the propositions.

Next, we aggregated the scores for each student answer proposition by taking the maximum average cosine across the different expected answer propositions. As in the previous experiment, the final score was the percentage of student answer propositions that achieved a score above the empirically-determined threshold. We tested thresholds between 0.05 and 0.95 in 0.05 increments. We measured the correlation between the LSA scores with the human ratings.

The best correlation was $r = 0.18$ (not significant), with the threshold at 0.10.[1] This is far below the performance of the previous approach which used LSA to compare entire sentences. Thus, these findings do not support hypothesis 1.

The decrease in the overall performance could potentially be due to the difference between comparing sentences (as in the original experiment) and comparing propositions. But the aggregate score essentially factors that out to the extent that length of string does not affect LSA discrimination. String length does affect LSA discrimination however. Rehder et al (1998) used LSA to assess the domain knowledge of essay writers. To determine the effect of essay length on LSA discrimination, they truncated each essay after 10 words, 20 words, and so on. Below 60 words, they found fairly poor performance. The performance steadily increased from there up to their 200 word maximum. Despite this finding, we have found performance approaching human abilities on our tutoring texts which have an average length of 16 words. Thus, we thought that any minor reduction in performance due to length would be offset by increased information provided by the pre-processing.

Analysis of cases of disagreement between LSA and the human raters showed that some items got very bad scores because one component consisted only of a "stop word" — a member of a list of 440 common words that includes prepositions, pronouns, and some very common adjectives, adverbs, verbs, and nouns. For example, one student proposition has a verb component group consisting of the string, "stores", and the expected answer has the verb string, "has". In this case ("RAM stores information being worked with"), the meanings of these two verbs are quite similar. But because "has" is on the stop word list, it has no representation in the LSA space, and the cosine comparison returns a value of 0.

On the other end of the spectrum, there was often an exact match between the subjects. For example, "RAM" and "CPU" are frequent subjects which, if they match at all, tend to match exactly, getting a 1.0 cosine. Because average "good" cosine matches are often in the 0.4 to 0.6 range, this tends to inflate the cosine average. This is especially the case for intransitive sentences where there are only two components. At the threshold that provided the best correlation with human raters, 0.10, the verb string only had to match at the 0.20 cosine level to put the entire proposition over threshold.

Another factor which seemed to affect the ratings was the fact that there are so different ways in which the same content can be expressed in natural language. For example, "RAM stores things being worked with" should have a fairly high semantic match for "The CPU uses RAM as a short-term memory storage" (whole string LSA cosine = 0.48). But because the components do not line up at all in this approach, the cosine average score is 0.03.

Based on these analyses, and under the hope that hypotheses 3 was the case instead of hypothesis 2, the approach was modified as described in the next section.

Experiment 2

As previously mentioned, the shortness of the subject components seemed to have an inordinate effect on the overall scores. The average number of words in subject components was 1.6, and many subject strings include stop words like "the" which do not contribute to LSA

[1] Due to the tediousness of pre-processing the sentences by hand, these results were only calculated on the first third of the test set. Analyses of the correlations on the original task on this part of the test set showed that it had lower performance ($r = 0.32, p = .01$), but not as low as the results of experiment 1. Immediate future work will be to process the rest of the test set.

cosines. Because of this, we tested in experiment two, an alternative scoring strategy. In this strategy, the score between two propositions was calculated as follows:

> If there is a suitable match between the subjects, then return the average of the cosines of the other components.

Here, "suitable match" was defined as either a cosine of 0.7[2], or a cosine of zero. In theory a zero cosine means a complete lack of semantic similarity. In practice, however, the cosine is only exactly 0 when one of the strings is empty modulo stop words. Thus, this allows the matching of vague subjects like "you".

There are psychological theories of discourse which (vaguely) support this approach. One is the Given-New distinction of referents in discourse (Clark & Haviland, 1977; Brennan, 1995). The theory includes a discourse processing strategy in which the hearer searches the prior discourse context for an antecedent for Given information which is commonly the syntactic subject of a sentence. The rest of the sentence is New information which is attached to the antecedent. In our approach, we filter out expected answers which do not have matching Given information. Then we rate the similarity with the remaining items based on the similarity of the New information.

For this approach, the results were better than for experiment 1. The maximum correlation between the system and the human raters was $r = 0.24, (p = 0.06)$. This still does not approach the level of performance of the original system, however. This led us to attempt to address the other concerns raised above in experiment 3.

Experiment 3

In experiment 3, we built on the Given-New approach presented above. This time, however, we joined the verb component of each proposition with its object component into one larger component. This corresponds to the VP in the basic S $\rightarrow$ NP VP sentence, or to the predicate in the Subject/Predicate description of a sentence. Obviously this is a partial reversal from our previous approach of adding more information derived from syntax. The justification was to make the LSA comparisons less brittle with respect to distinctions between information in the verb and in the object.

The results for this approach were better than for experiment 2. The maximum correlation was $r = 0.40 (p < 0.01)$, with a cosine threshold of 0.3. (The Cronbach's alpha score was $\alpha = 0.49$.)

Although this is an improvement, it is still not as good as the 0.49 correlation achieved by matching the entire sentence strings. Thus, these results do not support hypothesis 1. And taken together, their support for hypothesis 3 is ambiguous at best. This leaves us with the question: Why, when getting more information, does the discrimination still suffer?

[2]0.5 was also tested, but it made a negligible difference

Discussion and Future work

In some ways our approach has been to find the best formula for combining the similarity ratings between the different components. The one which worked best, the one used in experiment 3, is non-linear. Perhaps a further search of combination methods can out-perform the basic LSA approach.

Taking the cue from other statistical NLP approaches and neural networks, perhaps we just have to find the right weight space which gives the best correspondence between the parameters (components) and the training data (human judgments). Ideally, if we were to attempt such an implementation, instead of aggregate human judgments over a set of items, we would have a rating for each pair of items. That would be much more demanding on the human raters, but would give more data to train the approach on.

Future work will focus on two fronts. First, we will acquire more data on which to evaluate this approach, both by adding more test items, and by getting additional human judgments as outlined above. Second, we will explore other methods of combining the added syntactic-derived information into LSA.

Acknowledgments

This project was partially supported by grant number SBR 9720314 from the National Science Foundation's Learning and Intelligent Systems Unit. Many thanks to Mark Core for comments on this approach and to Katja Wiemer-Hastings for comments on the paper.

References

Abney, S. (1996). Partial parsing via finite-state cascades. In *Proceedings of the ESSLLI '96 Robust Parsing Workshop.*

Brennan, S. (1995). Centering attention in discourse. *Language and Cognitive Processes*, *10*, 137–167.

Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, *12*, 177–210.

Chomsky, N. (1981). Principles and parameters in syntactic theory. In Hornstein, N., & Lightfoot, D. (Eds.), *Explanation in Linguistics: The Logical Problem of Language Acquisition*. Longman, London.

Christiansen, M., & Chater, N. (1999). Connectionist Natural Language Processing: the state of the art. *Cognitive Science*, *23*(4), 417–437.

Clark, H. H., & Haviland, S. E. (1977). Comprehension and the given-new contract. In Freedle, R. (Ed.), *Discourse production and comprehension*, pp. 1–40. Earlbaum, Hillsdale NJ.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Finch, S., & Chater, N. (1992). Bootstrapping syntactic categories using unsupervised learning. In *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society*, pp. 820–825 Hillsdale, NJ. Lawrence Erlbaum Associates Inc.

Gleitman, L., & Gillette, J. (1994). The role of syntax in verb learning. In Fletcher, P., & MacWhinney, B. (Eds.), *The Handbook of Child Language*. Blackwell, Oxford UK.

Graesser, A. C., Franklin, S. P., Wiemer-Hastings, P., & the Tutoring Research Group (1998). Simulating smooth tutorial dialogue with pedagogical value. In *Proceedings of the 11th International Florida Artificial Intelligence Research Symposium Conference*, pp. 163–167. AAAI Press.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press, Cambridge, MA.

Landauer, T. K., Laham, D., Rehder, R., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pp. 412–417 Mahwah, NJ. Erlbaum.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

McDonald, S. (2000). *Environmental determinants of lexical processing effort*. Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland.

Packiam-Alloway, T., Ramscar, M., & Corley, M. (1999). Verbal versus embodied priming in schema mapping tasks. In *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society*. Laurence Earlbaum Assocs. Vancouver, Canada, August, 1999.

Person, N. K., Graesser, A. C., Magliano, J. P., & Kreuz, R. J. (1994). Inferring what the student knows in one-to-one tutoring: The role of student questions and answers. *Learning and Individual Differences*, *6*, 205–229.

Rehder, B., Schreiner, M., Laham, D., Wolfe, M., Landauer, T., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes*, *25*, 337–354.

Schank, R., & Riesbeck, C. (Eds.). (1981). *Inside computer understanding*. Erlbaum, Hillsdale, NJ.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39–91.

Wiemer-Hastings, P., Graesser, A., Harter, D., & the Tutoring Research Group (1998). The foundations and architecture of AutoTutor. In Goettl, B., Halff, H., Redfield, C., & Shute, V. (Eds.), *Intelligent Tutoring Systems, Proceedings of the 4th International Conference*, pp. 334–343 Berlin. Springer.

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In Lajoie, S., & Vivet, M. (Eds.), *Artificial Intelligence in Education*, pp. 535–542 Amsterdam. IOS Press.