Empirical Measurements of Lexical Similarity in Noun Phrase Conjuncts

Deirdre Hogan*
Department of Computer Science
Trinity College Dublin
Dublin 2, Ireland
dhogan@computing.dcu.ie

Abstract

The ability to detect similarity in conjunct heads is potentially a useful tool in helping to disambiguate coordination structures - a difficult task for parsers. We propose a distributional measure of similarity designed for such a task. We then compare several different measures of word similarity by testing whether they can empirically detect similarity in the head nouns of noun phrase conjuncts in the Wall Street Journal (WSJ) treebank. We demonstrate that several measures of word similarity can successfully detect conjunct head similarity and suggest that the measure proposed in this paper is the most appropriate for this task.

1 Introduction

Some noun pairs are more likely to be conjoined than others. Take the follow two alternate bracketings: 1. *busloads of ((executives) and (their spouses))* and 2. *((busloads of executives) and (their spouses))*. The two head nouns coordinated in 1 are *executives* and *spouses*, and (incorrectly) in 2: *busloads* and *spouses*. Clearly, the former pair of head nouns is more likely and, for the purpose of discrimination, a parsing model would benefit if it could learn that *executives and spouses* is a more likely combination than *busloads and spouses*. If nouns co-occurring in coordination patterns are often semantically similar, and if a similarity measure could be defined so that, for example: $sim(executives, spouses) > sim(busloads, spouses)$ then it is potentially useful for coordination disambiguation.

The idea that nouns co-occurring in conjunctions tend to be semantically related has been noted in (Riloff and Shepherd, 1997) and used effectively to automatically cluster semantically similar words (Roark and Charniak, 1998; Caraballo, 1999; Widdows and Dorow, 2002). The tendency for conjoined nouns to be semantically similar has also been exploited for coordinate noun phrase disambiguation by Resnik (1999) who employed a measure of similarity based on WordNet to measure which were the head nouns being conjoined in certain types of coordinate noun phrase.

In this paper we look at different measures of word similarity in order to discover whether they can detect empirically a tendency for conjoined nouns to be more similar than nouns which co-occur but are not conjoined. In Section 2 we introduce a measure of word similarity based on word vectors and in Section 3 we briefly describe some WordNet similarity measures which, in addition to our word vector measure, will be tested in the experiments of Section 4.

2 Similarity based on Coordination Co-occurrences

The potential usefulness of a similarity measure depends on the particular application. An obvious place to start, when looking at similarity functions for measuring the type of semantic similarity common for coordinate nouns, is a similarity function based on distributional similarity with context de-

* Now at the National Centre for Language Technology, Dublin City University, Ireland.

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 149–152,
Prague, June 2007.

fined in terms of coordination patterns. Our measure of similarity is based on noun co-occurrence information, extracted from conjunctions and lists. We collected co-occurrence data on $82,579$ distinct word types from the BNC and the WSJ treebank.

We extracted all noun pairs from the BNC which occurred in a pattern of the form: *noun cc noun*[1], as well as lists of any number of nouns separated by commas and ending in *cc noun*. Each noun in the list is linked with every other noun in the list. Thus for a list: n_1, n_2, and n_3, there will be co-occurrences between words n_1 and n_2, between n_1 and n_3 and between n_2 and n_3. To the BNC data we added all head noun pairs from the WSJ (sections 02 to 21) that occurred together in a coordinate noun phrase.[2]

From the co-occurrence data we constructed word vectors. Every dimension of a word vector represents another word type and the values of the components of the vector, the term weights, are derived from the coordinate word co-occurrence counts. We used dampened co-occurrence counts, of the form: $1 + log(count)$, as the term weights for the word vectors. To measure the similarity of two words, w_1 and w_2, we calculate the cosine of the angle between the two word vectors, $\vec{w}_1$ and $\vec{w}_2$.

3 WordNet-Based Similarity Measures

We also examine the following measures of semantic similarity which are WordNet-based.[3] Wu and Palmer (1994) propose a measure of similarity of two concepts c_1 and c_2 based on the depth of concepts in the WordNet hierarchy. Similarity is measured from the depth of the most specific node dominating both c_1 and c_2, (their lowest common subsumer), and normalised by the depths of c_1 and c_2. In (Resnik, 1995) concepts in WordNet are augmented by corpus statistics and an information-theoretic measure of semantic similarity is calculated. Similarity of two concepts is measured by the information content of their lowest common subsumer in the *is-a* hierarchy of WordNet. Both Jiang and Conrath (1997) and Lin (1998) propose extentions of Resnik's measure. Leacock and Chodorow (1998)'s measure takes into account the path length between two concepts, which is scaled by the depth of the hierarchy in which they reside. In (Hirst and St-Onge, 1998) similarity is based on path length as well as the number of changes in the direction in the path. In (Banerjee and Pedersen, 2003) semantic relatedness between two concepts is based on the number of shared words in their WordNet definitions (glosses). The gloss of a particular concept is extended to include the glosses of other concepts to which it is related in the WordNet hierarchy. Finally, Patwardhan and Pederson (2006) build on previous work on second-order co-occurrence vectors (Schütze, 1998) by constructing second-order co-occurrence vectors from WordNet glosses, where, as in (Banerjee and Pedersen, 2003), the gloss of a concept is extended so that it includes the gloss of concepts to which it is directly related in WordNet.

4 Experiments

We selected two sets of data from sections 00, 01, 22 and 24 of the WSJ treebank. The first consists of all nouns pairs which make up the head words of two conjuncts in coordinate noun phrases (again not including coordinate NPBs). We found 601 such coordinate noun pairs. The second data set consists of 601 word pairs which were selected at random from all head-modifier pairs where both head and modifier words are nouns and are *not* coordinated. We tested the 9 different measures of word similarity just described on each data set in order to see if a significant difference could be detected between the similarity scores for the coordinate words sample and non-coordinate words sample.

Initially both the coordinate and non-coordinate pair samples each contained 601 word pairs. However, before running the experiments we removed all pairs where the words in the pair were identical. This is because identical words occur more often in coordinate head words than in other lexical dependencies (there were 43 pairs with identical words in the coordination set, compared to 3 such pairs in the

[1] It would be preferable to ensure that the pairs extracted are unambiguously conjoined heads. We leave this to future work.

[2] We did not include coordinate head nouns from base noun phrases (NPB) (i.e. noun phrases that do not dominate other noun phrases) because the underspecified annotation of NPBs in the WSJ means that the conjoined head nouns can not always be easily identified.

[3] All of the WordNet-based similarity measure experiments, as well as a random similarity measure, were carried out with the WordNet::Similarity package, http://search.cpan.org/dist/WordNet-Similarity.

SimTest	n_{coord}	$\overline{x}_{coord}$	SD_{coord}	$n_{nonCoord}$	$\overline{x}_{nonCoord}$	$SD_{nonCoord}$	95% CI	p-value
coordDistrib	503	0.11	0.13	485	0.06	0.09	[0.04 0.07]	0.000
(Resnik, 1995)	444	3.19	2.33	396	2.43	2.10	[0.46 1.06]	0.000
(Lin, 1998)	444	0.27	0.26	396	0.19	0.22	[0.04 0.11]	0.000
(Jiang and Conrath, 1997)	444	0.13	0.65	395	0.07	0.08	[-0.01 0.11]	0.083
(Wu and Palmer, 1994)	444	0.63	0.19	396	0.55	0.19	[0.06 0.11]	0.000
(Leacock and Chodorow, 1998)	444	1.72	0.51	396	1.52	0.47	[0.13 0.27]	0.000
(Hirst and St-Onge, 1998)	459	1.599	2.03	447	1.09	1.87	[0.25 0.76]	0.000
(Banerjee and Pedersen, 2003)	451	114.12	317.18	436	82.20	168.21	[-1.08 64.92]	0.058
(Patwardhan and Pedersen, 2006)	459	0.67	0.18	447	0.66	0.2	[-0.02 0.03]	0.545
random	483	0.89	0.17	447	0.88	0.18	[-0.02 0.02]	0.859

Table 1: Summary statistics for 9 different word similarity measures (plus one random measure):n_{coord} and $n_{nonCoord}$ are the sample sizes for the coordinate and non-coordinate noun pairs samples, respectively; $\overline{x}_{coord}$, SD_{coord} and $\overline{x}_{nonCoord}$, $SD_{nonCoord}$ are the sample means and standard deviations for the two sets. The 95% CI column shows the 95% confidence interval for the difference between the two sample means. The p-value is for a Welch two sample two-sided t-test. *coordDistrib* is the measure introduced in Section 2.

non-coordination set). If we had not removed them, a statistically significant difference between the similarity scores of the pairs in the two sets could be found simply by using a measure which, say, gave one score for identical words and another (lower) score for all non-identical word pairs.

Results for all similarity measure tests on the data sets described above are displayed in Table 1. In one final experiment we used a random measure of similarity. For each experiment we produced two samples, one consisting of the similarity scores given by the similarity measure for the coordinate noun pairs, and another set of similarity scores generated for the non-coordinate pairs. The sample sizes, means, and standard deviations for each experiment are shown in the table. Note that the variation in the sample size is due to coverage: the different measures did not produce a score for all word pairs. Also displayed in Table 1 are the results of statistical significance tests based on the Welsh two sample t-test. A 95% confidence interval for the difference of the sample means is shown along with the p-value.

5 Discussion

For all but three of the experiments (excluding the random measure), the difference between the mean similarity measures is statistically significant. Interestingly, the three tests where no significant difference was measured between the scores on the coordination set and the non-coordination set (Jiang and Conrath, 1997; Banerjee and Pedersen, 2003; Patwardhan and Pedersen, 2006) were the three top scoring measures in (Patwardhan and Pedersen, 2006), where a subset of six of the above WordNet-based experiments were compared and the measures evaluated against human relatedness judgements and in a word sense disambiguation task. In another comparative study (Budanitsky and Hirst, 2002) of five of the above WordNet-based measures, evaluated as part of a real-word spelling correction system, Jiang and Conrath (1997)'s similarity score performed best. Although performing relatively well under other evaluation criteria, these three measures seem less suited to measuring the kind of similarity occurring in coordinate noun pairs. One possible explanation for the unsuitability of the measures of (Patwardhan and Pedersen, 2006) for the coordinate similarity task could be based on how context is defined when building context vectors. Context for an instance of the the word *w* is taken to be the words that surround *w* in the corpus within a given number of positions, where the corpus is taken as all the glosses in WordNet. Words that form part of collocations such as *disk drives* or *task force* would then tend to have very similar contexts, and thus such word pairs, from non-coordinate modifier-head relations, could be given too high a similarity score.

Although the difference between the mean similarity scores seems rather slight in all experiments, it is worth noting that not all coordinate head words *are* semantically related. To take a couple of examples from the coordinate word pair set: *work/harmony* extracted from *hard work and harmony*, and *power/clause* extracted from *executive power and the appropriations clause*. We would not expect these word pairs to get a high similarity score. On the other hand, it is also possible that some of the examples of non-coordinate dependencies involve semantically similar words. For example, nouns in lists are often semantically similar, and we did not exclude nouns extracted from lists from the non-coordinate test set.

Although not all coordinate noun pairs are semantically similar, it seems clear, on inspection of the two sets of data, that they are more likely to be semantically similar than modifier-head word pairs, and the tests carried out for most of the measures of semantic similarity detect a significant difference between the similarity scores assigned to coordinate pairs and those assigned to non-coordinate pairs.

It is not possible to judge, based on the significance tests alone, which might be the most useful measure for the purpose of disambiguation. However, in terms of coverage, the distributional measure introduced in Section 2 clearly performs best[4]. This measure of distributional similarity is perhaps more suited to the task of coordination disambiguation because it directly measures the type of similarity that occurs between coordinate nouns. That is, the distributional similarity measure presented in Section 2 defines two words as similar if they occur in coordination patterns with a similar set of words and with similar distributions. Whether the words are *semantically* similar becomes irrelevant. A measure of semantic similarity, on the other hand, might find words similar which are quite unlikely to appear in coordination patterns. For example, Cederberg and Widdows (2003) note that words appearing in coordination patterns tend to be on the same ontological level: ‘fruit and vegetables’ is quite likely to occur, whereas ‘fruit and apples’ is an unlikely co-occurrence. A WordNet-based measure of semantic similarity, however, might give a high score to both of the noun pairs.

In the future we intend to use the similarity measure outlined in Section 2 in a lexicalised parser to help resolve coordinate noun phrase ambiguities.

Acknowledgements Thanks to the TCD Broad Curriculum Fellowship and to the SFI Research Grant 04/BR/CS370 for funding this research. Thanks also to Pádraig Cunningham, Saturnino Luz and Jennifer Foster for helpful discussions.

[4]Somewhat unsurprisingly given it is part trained on data from the same domain.

References

Satanjeev Banerjee and Ted Pedersen. 2003 Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *Proceeding of the 18th IJCAI.*

Alexander Budanitsky and Graeme Hirst. 2002 Semantic Distance in WordNet: An experimental, application-oriented Evaluation of Five Measures In *Proceedings of the 3rd CICLING.*

Sharon Caraballo. 1999 Automatic construction of a hypernym-labeled noun hierarchy from text In *Proceedings of the 37th ACL.*

Scott Cederberg and Dominic Widdows. 2003. Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. In *Proceedings of the 7th CoNLL.*

G. Hirst and D. St-Onge 1998. Lexical Chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*. MIT Press.

J. Jiang and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the ROCLING*.

C. Leacock and M. Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*. MIT Press.

D. Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th ICML.*

Siddharth Patwardhan and Ted Pedersen. 2006. Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proceedings of Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together, EACL.*

Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity. In *Proceedings of IJCAI.*

Philip Resnik. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. In *Journal of Artificial Intelligence Research*, 11:95-130.

Ellen Riloff and Jessica Shepherd 1997. A Corpus-based Approach for Building Semantic Lexicon. In *Proceedings of the 2nd EMNLP*.

Brian Roark and Eugene Charniak 1998. Noun-phrase Co-occurrence Statistics for Semi-automatic semantic lexicon construction. In *Proceedings of the COLING-ACL.*

Hinrich Schütze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97-123.

Dominic Widdows and Beate Dorow. 2002. A Graph Model for Unsupervised Lexical Acquisition. In *Proceedings of the 19th COLING.*

Zhibiao Wu and Martha Palmer. 1994. Verb Semantics and Lexical Selection. In *Proceedings of the ACL.*