Proceedings of the Australasian Language Technology Workshop 2007, pages 146-148

Distributional Similarity of Multi-Word Expressions

Laura Ingram and **James R. Curran**
School of Information Technologies
The University of Sydney
NSW 2006, Australia
{ling6188, james}@it.usyd.edu.au

Abstract

Most existing systems for automatically extracting lexical-semantic resources neglect multi-word expressions (MWEs), even though approximately 30% of gold-standard thesauri entries are MWEs.

We present a distributional similarity system that identifies synonyms for MWEs. We extend Grefenstette's SEXTANT shallow parser to first identify bigram MWEs using collocation statistics from the Google WEB1T corpus. We extract contexts from WEB1T to increase coverage on the sparser bigrams.

1 Introduction

Lexical-semantic resources, such as WordNet (Fellbaum, 1998), are used in many applications in Natural Language Processing (NLP). Unfortunately, they are expensive and time-consuming to produce and are prone to bias and limited coverage. Automatically extracting these resources is crucial to overcoming the knowledge bottleneck in NLP.

Existing distributional approaches to semantic similarity focus on unigrams, with very little work on extracting synonyms for multi-word expressions (MWEs). In this work, we extend an existing system to support MWEs by identifying bigram MWEs using collocation statistics (Manning and Schütze, 1999). These are calculated using n-gram counts from the Google WEB1T corpus (Brants and Franz, 2006).

We evaluate against several gold-standard thesauri and observe a slight decrease in overall performance when the bigram MWEs were included. This is unsurprising since the larger vocabulary and sparser contextual information for bigrams makes the task significantly harder. We also experimented with contexts extracted from WEB1T in an attempt to overcome the data sparseness problem. Inspection of the results for individual headwords revealed many cases where the synonyms returned were significantly better when bigram data was included.

2 Background

Distributional similarity relies on the *distributional hypothesis* that similar terms appear in similar contexts (Harris, 1954). Here we extend the SEXTANT parser (Grefenstette, 1994) to include multi-word *terms* and syntactic *contexts*.

Curran (2004) experiments with different parsers for extracting contextual information, including SEXTANT, MINIPAR (Lin, 1994), RASP (Briscoe and Carroll, 2002), and CASS (Abney, 1996). Lin (1998) used MINIPAR and Weeds (2003) used RASP for distributional similarity calculations. MINIPAR is the only parser to identify a range of MWEs that has been used for distributional similarity. Weeds (2003) and Curran (2004) evaluate measures for calculating distributional similarity. We follow (Curran, 2004) in using the weighted Jaccard measure with truncated t-test relation weighting for our experiments.

3 Detecting MWEs

The initial step in creating a thesaurus for MWEs is to identify potential MWE headwords using collocation statistics. We used various statistical tests, e.g. the t-test and the log-likelihood test (Manning and Schütze, 1999), calculated over the Google WEB1T unigram and bigram counts. These counts, calculated over 1 trillion words of web text, gave the most reliable counts. However, highly ranked terms, e.g.

Contact Us and Site Map, demonstrate bias towards web-related terminology. This list of selected bigrams is used to detect bigrams within the BNC using a modified version of the Viterbi algorithm.

4 Context Extraction

Grefenstette's (2004) (SEXTANT) parser was extended to extract contextual information for the list of selected bigrams extracted above. Adding these bigrams does not result in a substantial increase in the number of relations which implies that there is very little contextual information available about the bigram data. This has a significant impact on the difficulty of the task.

Experiments were also conducted whereby the contextual information was extracted from the WEB1T 3, 4 and 5-gram data for a list of known bigrams from the gold-standard thesauri. This data lacks the syntactic information provided by SEXTANT but the counts are estimated over 10,000 times as much data. This should reduce the sparseness problem for the bigram headwords.

5 Synonym Extraction

Following Curran (2004), the extracted synonyms are compared directly against multiple gold-standard thesauri. We extend this evaluation to include multi-word headwords and synonyms. We randomly selected 300 unigram and 300 bigram headwords from the MAQCUARIE (Bernard, 1990), MOBY (Ward, 1996), and ROGET'S (1911) thesauri, and WORDNET (Fellbaum, 1998).

We calculated the number of direct matches against the gold standard (DIRECT) and the inverse rank (INVR), the sum of the reciprocal ranks of matches. The results for the unigram headword experiments are summarised in Table 1.

Both INVR and DIRECT demonstrate that performance decreases when MWEs are included. However, performance did increase significantly for some terms when MWEs were added. For example, tool improved from 0.270 to 0.568 INVR. The results for rate, shown in Table 2, also improved.

The next set of experiments extracted synonyms for 300 bigram headwords drawn from the MACQUARIE thesaurus. The best results for bigram headwords was achieved when unigram and bigram data was extracted from WEB1T and the VPC resource (Baldwin and Villavicencio, 2002) was included. Table 3 shows the top 5 synonyms (as ranked by the Jaccard measure) for atomic bomb and dining table.

		DIRECT	INVR
BNC	UNI	22.6	1.717
t-test	UNI+ BI	22.2	1.650
	UNI+ BI+ VPC	22.2	1.659
WEB1T	3UNI	16.9	1.182
t-test	4UNI	19.3	1.454
	3UNI+ 4BI	15.6	1.004
	4UNI+ 5BI	19.8	1.344
	3UNI+ 4BI+ 4VPC	15.6	1.001
	4UNI+ 5BI+ 5VPC	19.8	1.346
WEB1T	3UNI+ 4BI	17.5	1.185
THES	4UNI+ 5BI	17.5	1.194
	3UNI+ 4BI+ 4VPC	17.5	1.187
	4UNI+ 5BI+ 5VPC	21.2	1.491

Table 1: Results for unigram headwords

UNI	UNI+ BI	UNI+ BI+ VPC
level	level	level
price	price	price
cost	amount	cost
income	cost	amount
growth	speed	average

Table 2: Sample synonyms for rate

ATOMIC BOMB	DINING TABLE
nuclear bomb	coffee table
atom bomb	dining room
nuclear explosion	cocktail table
atomic explosion	dining chair
nuclear weapon	bedroom furniture

Table 3: Sample bigram synonyms

6 Conclusion

We have integrated the identification of simple multi-word expressions (MWEs) with a state-of-the-art distributional similarity system. We evaluated extracted synonyms for both unigram and bigram headwords against a gold standard consisting of the union of multiple thesauri.

The main difficulties are the sparsity of distributional evidence for MWEs and their low coverage in the gold standard. These preliminary experiments show the potential of distributional similarity for extracting lexical-semantic resources for both unigrams and MWEs.

References

Steven Abney. 1996. Partial parsing via finite-state cascades. *Journal of Natural Language Engineering*, 2(4):337–344.

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the Sixth Conference on Computational Natural Language Learning (CoNLL 2002)*, pages 98–104, Taipei, Taiwan.

John R.L. Bernard, editor. 1990. *The Macquarie Encyclopedic Thesaurus*. The Macquarie Library, Sydney, Australia.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram corpus version 1. Technical report, Google Inc.

Ted Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1499–1504, Las Palmas de Gran Canaria, 29-31 May.

James R. Curran. 2004. *From Distributional to Semantic Similarity*. Ph.D. thesis, University of Edinburgh.

Chrisriane Fellbaum, editor. 1998. *Wordnet: An Electronic Lexical Database*. MIT Press, Cambridge.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, USA.

Zellig Harris. 1954. Distributional structure. *Word*, 10(2/3):146–162.

Dekang Lin. 1994. Principar – an efficient, broad-coverage, principle-based parser. In *Proceedings of COLING-94*, pages 482–488, Kyoto, Japan.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume 2, pages 768–774, Montreal, Quebec, Canada.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Language Processing*. MIT Press, Cambridge, MA USA.

Peter Mark Roget. 1911. *Thesaurus of English Words and Phrases*. Longmans, Green and Company, London, UK.

Grady Ward. 1996. Moby thesaurus. http://etext.icewire.com/moby/.

Julie E. Weeds. 2003. *Measures and Applications of Lexical Distributional Similarity*. Ph.D. thesis, University of Sussex.