The Distributional Inclusion Hypotheses and Lexical Entailment

Maayan Geffet
School of Computer Science and Engineering
Hebrew University, Jerusalem, Israel, 91904
mary@cs.huji.ac.il

Ido Dagan
Department of Computer Science
Bar-Ilan University, Ramat-Gan, Israel, 52900
dagan@cs.biu.ac.il

Abstract

This paper suggests refinements for the Distributional Similarity Hypothesis. Our proposed hypotheses relate the distributional behavior of pairs of words to *lexical entailment* – a tighter notion of semantic similarity that is required by many NLP applications. To automatically explore the validity of the defined hypotheses we developed an *inclusion testing algorithm* for characteristic features of two words, which incorporates corpus and web-based feature sampling to overcome data sparseness. The degree of hypotheses validity was then empirically tested and manually analyzed with respect to the word sense level. In addition, the above testing algorithm was exploited to improve lexical entailment acquisition.

1 Introduction

Distributional Similarity between words has been an active research area for more than a decade. It is based on the general idea of Harris' Distributional Hypothesis, suggesting that words that occur within similar contexts are semantically similar (Harris, 1968). Concrete similarity measures compare a pair of weighted context feature vectors that characterize two words (Church and Hanks, 1990; Ruge, 1992; Pereira et al., 1993; Grefenstette, 1994; Lee, 1997; Lin, 1998; Pantel and Lin, 2002; Weeds and Weir, 2003).

As it turns out, distributional similarity captures a somewhat loose notion of semantic similarity (see Table 1). It does not ensure that the meaning of one word is *preserved* when replacing it with the other one in some context.

However, many semantic information-oriented applications like Question Answering, Information Extraction and Paraphrase Acquisition require a tighter similarity criterion, as was also demonstrated by papers at the recent PASCAL Challenge on Recognizing Textual Entailment (Dagan et al., 2005). In particular, all these applications need to know when the meaning of one word can be inferred (entailed) from another word, so that one word could substitute the other in some contexts. This relation corresponds to several lexical semantic relations, such as synonymy, hyponymy and some cases of meronymy. For example, in Question Answering, the word *company* in a question can be substituted in the text by *firm* (synonym), *automaker* (hyponym) or *division* (meronym). Unfortunately, existing manually constructed resources of lexical semantic relations, such as WordNet, are not exhaustive and comprehensive enough for a variety of domains and thus are not sufficient as a sole resource for application needs[1].

Most works that attempt to learn such concrete lexical semantic relations employ a co-occurrence pattern-based approach (Hearst, 1992; Ravichandran and Hovy, 2002; Moldovan et al., 2004). Typically, they use a set of predefined lexico-syntactic patterns that characterize specific semantic relations. If a candidate word pair (like *company-automaker*) co-occurs within the same sentence satisfying a concrete pattern (like " …companies, *such as* automakers"), then it is expected that the corresponding semantic relation holds between these words (hypernym-hyponym in this example).

In recent work (Geffet and Dagan, 2004) we explored the correspondence between the distributional characterization of two words (which may hardly co-occur, as is usually the case for syno-

[1] We found that less than 20% of the lexical entailment relations extracted by our method appeared as direct or indirect WordNet relations (synonyms, hyponyms or meronyms).

Proceedings of the 43rd Annual Meeting of the ACL, pages 107–114,
Ann Arbor, June 2005.

<=> element, component	**<=> gap, spread**	*** town, airport**	**<= loan, mortgage**
=> government, body	*** warplane, bomb**	**<=> program, plan**	*** tank, warplane**
*** match, winner**	**=> bill, program**	**<= conflict, war**	**=> town, location**

Table 1: Sample of the data set of top-40 distributionally similar word pairs produced by the *RFF*-based method of (Geffet and Dagan, 2004). Entailment judgments are marked by the arrow direction, with '*' denoting no entailment.

nyms) and the kind of tight semantic relationship that might hold between them. We formulated a lexical entailment relation that corresponds to the above mentioned substitutability criterion, and is termed *meaning entailing substitutability* (which we term here for brevity as *lexical entailment*). Given a pair of words, this relation holds if there are some contexts in which one of the words can be substituted by the other, such that the meaning of the original word can be inferred from the new one. We then proposed a new feature weighting function (*RFF*) that yields more accurate distributional similarity lists, which better approximate the lexical entailment relation. Yet, this method still applies a standard measure for distributional vector similarity (over vectors with the improved feature weights), and thus produces many loose similarities that do not correspond to entailment.

This paper explores more deeply the relationship between distributional characterization of words and lexical entailment, proposing two new hypotheses as a refinement of the distributional similarity hypothesis. The main idea is that if one word entails the other then we would expect that virtually all the characteristic context features of the entailing word will actually occur also with the entailed word.

To test this idea we developed an automatic method for testing feature inclusion between a pair of words. This algorithm combines corpus statistics with a web-based feature sampling technique. The web is utilized to overcome the data sparseness problem, so that features which are not found with one of the two words can be considered as truly distinguishing evidence.

Using the above algorithm we first tested the empirical validity of the hypotheses. Then, we demonstrated how the hypotheses can be leveraged in practice to improve the precision of automatic acquisition of the entailment relation.

2 Background

2.1 Implementations of Distributional Similarity

This subsection reviews the relevant details of earlier methods that were utilized within this paper.

In the computational setting contexts of words are represented by feature vectors. Each word *w* is represented by a feature vector, where an entry in the vector corresponds to a feature *f*. Each feature represents another word (or term) with which *w* co-occurs, and possibly specifies also the syntactic relation between the two words as in (Grefenstette, 1994; Lin, 1998; Weeds and Weir, 2003). Pado and Lapata (2003) demonstrated that using syntactic dependency-based vector space models can help distinguish among classes of different lexical relations, which seems to be more difficult for traditional "bag of words" co-occurrence-based models.

A syntactic feature is defined as a triple *<term, syntactic_relation, relation_direction>* (the direction is set to 1, if the feature is the word's modifier and to 0 otherwise). For example, given the word "company" the feature *<earnings_report, gen, 0>* (genitive) corresponds to the phrase "company's earnings report", and *<profit, pcomp, 0>* (prepositional complement) corresponds to "the profit of the company". Throughout this paper we used syntactic features generated by the Minipar dependency parser (Lin, 1993).

The value of each entry in the feature vector is determined by some weight function *weight(w,f)*, which quantifies the degree of statistical association between the feature and the corresponding word. The most widely used association weight function is (point-wise) Mutual Information (*MI*) (Church and Hanks, 1990; Lin, 1998; Dagan, 2000; Weeds et al., 2004).

Once feature vectors have been constructed, the similarity between two words is defined by some vector similarity metric. Different metrics have been used, such as weighted Jaccard (Grefenstette, 1994; Dagan, 2000), cosine (Ruge, 1992), various information theoretic measures (Lee, 1997), and the widely cited and competitive (see (Weeds and Weir, 2003)) measure of Lin (1998) for similarity between two words, *w* and *v*, defined as follows:

$$sim_{Lin}(w,v) = \frac{\sum_{f \in F(w) \cap F(v)} weight(w,f) + weight(v,f)}{\sum_{f \in F(w)} weight(w,f) + \sum_{f \in F(v)} weight(v,f)},$$

where *F(w)* and *F(v)* are the active features of the two words (positive feature weight) and the weight function is defined as *MI*. As typical for vector similarity measures, it assigns high similarity scores if many of the two word's features overlap, even though some prominent features might be disjoint. This is a major reason for getting such semantically loose similarities, like *company - government* and *country - economy*.

Investigating the output of Lin's (1998) similarity measure with respect to the above criterion in (Geffet and Dagan, 2004), we discovered that the quality of similarity scores is often hurt by inaccurate feature weights, which yield rather noisy feature vectors. Hence, we tried to improve the feature weighting function to promote those features that are most indicative of the word meaning. A new weighting scheme was defined for bootstrapping feature weights, termed *RFF* (Relative Feature Focus). First, basic similarities are generated by Lin's measure. Then, feature weights are recalculated, boosting the weights of features that characterize many of the words that are most similar to the given one[2]. As a result the most prominent features of a word are concentrated within the top-100 entries of the vector. Finally, word similarities are recalculated by Lin's metric over the vectors with the new *RFF* weights.

The lexical entailment prediction task of (Geffet and Dagan, 2004) measures how many of the top ranking similarity pairs produced by the *RFF*-based metric hold the entailment relation, in at least one direction. To this end a data set of 1,200 pairs was created, consisting of top-*N (N=40)* similar words of 30 randomly selected nouns, which were manually judged by the lexical entailment criterion. Quite high Kappa agreement values of 0.75 and 0.83 were reported, indicating that the entailment judgment task was reasonably well defined. A subset of the data set is demonstrated in Table 1.

The *RFF* weighting produced 10% precision improvement over Lin's original use of MI, suggesting the *RFF* capability to promote semantically meaningful features. However, over 47% of the word pairs in the top-40 similarities are not related by entailment, which calls for further improvement. In this paper we use the same data set [3] and the *RFF* metric as a basis for our experiments.

2.2 Predicting Semantic Inclusion

Weeds et al. (2004) attempted to refine the distributional similarity goal to predict whether one term is a generalization/specification of the other. They present a *distributional generality* concept and expect it to correlate with semantic generality. Their conjecture is that the majority of the features of the more specific word are included in the features of the more general one. They define the feature *recall* of *w* with respect to *v* as the weighted proportion of features of *v* that also appear in the vector of *w*. Then, they suggest that a hypernym would have a higher feature recall for its hyponyms (specifications), than vice versa.

However, their results in predicting the hyponymy-hyperonymy direction (71% precision) are comparable to the naïve baseline (70% precision) that simply assumes that general words are more frequent than specific ones. Possible sources of noise in their experiment could be ignoring word polysemy and data sparseness of word-feature co-occurrence in the corpus.

3 The Distributional Inclusion Hypotheses

In this paper we suggest refined versions of the distributional similarity hypothesis which relate distributional behavior with lexical entailment.

[2] In concrete terms *RFF* is defined by:

$RFF(w,f) = \sum_{v \in WS(f) \cap N(w)} sim(w,v)$,

where *sim(w,v)* is an initial approximation of the similarity space by Lin's measure, *WS(f)* is a set of words co-occurring with feature *f*, and *N(w)* is the set of the most similar words of *w* by Lin's measure.

3 Since the original data set did not include the direction of entailment, we have enriched it by adding the judgments of entailment direction.

Extending the rationale of Weeds et al., we suggest that if the meaning of a word *v* entails another word *w* then it is expected that all the typical contexts (features) of *v* will occur also with *w*. That is, the characteristic contexts of *v* are expected to be included within all *w*'s contexts (but not necessarily amongst the most characteristic ones for *w*). Conversely, we might expect that if *v*'s characteristic contexts are included within all *w*'s contexts then it is likely that the meaning of *v* does entail *w*. Taking both directions together, lexical entailment is expected to highly correlate with characteristic feature inclusion.

Two additional observations are needed before concretely formulating these hypotheses. As explained in Section 2, word contexts should be represented by syntactic features, which are more restrictive and thus better reflect the restrained semantic meaning of the word (it is difficult to tie entailment to looser context representations, such as co-occurrence in a text window). We also notice that distributional similarity principles are intended to hold at the sense level rather than the word level, since different senses have different characteristic contexts (even though computational common practice is to work at the word level, due to the lack of robust sense annotation).

We can now define the two *distributional inclusion hypotheses*, which correspond to the two directions of inference relating distributional feature inclusion and lexical entailment. Let v_i and w_j be two word senses of the words *w* and *v*, correspondingly, and let $v_i => w_j$ denote the (directional) entailment relation between these senses. Assume further that we have a measure that determines the set of *characteristic* features for the meaning of each word sense. Then we would hypothesize:

Hypothesis I:

If $v_i => w_j$ then all the characteristic (syntactic-based) features of v_i are expected to appear with w_j.

Hypothesis II:

If all the characteristic (syntactic-based) features of v_i appear with w_j then we expect that $v_i => w_j$.

4 Word Level Testing of Feature Inclusion

To check the validity of the hypotheses we need to test feature inclusion. In this section we present an automated word-level feature inclusion testing method, termed ITA (*Inclusion Testing Algorithm*). To overcome the data sparseness problem we incorporated web-based feature sampling. Given a test pair of words, three main steps are performed, as detailed in the following subsections:

Step 1: Computing the set of characteristic features for each word.

Step 2: Testing feature inclusion for each pair, in both directions, within the given corpus data.

Step 3: Complementary testing of feature inclusion for each pair in the web.

4.1 Step 1: Corpus-based generation of characteristic features

To implement the first step of the algorithm, the *RFF* weighting function is exploited and its top-100 weighted features are taken as most characteristic for each word. As mentioned in Section 2, (Geffet and Dagan, 2004) shows that *RFF* yields high concentration of good features at the top of the vector.

4.2 Step 2: Corpus-based feature inclusion test

We first check feature inclusion in the corpus that was used to generate the characteristic feature sets. For each word pair *(w, v)* we first determine which features of *w* do co-occur with *v* in the corpus. The same is done to identify features of *v* that co-occur with *w* in the corpus.

4.3 Step 3: Complementary Web-based Inclusion Test

This step is most important to avoid inclusion misses due to the data sparseness of the corpus. A few recent works (Ravichandran and Hovy, 2002; Keller et al., 2002; Chklovski and Pantel, 2004) used the web to collect statistics on word co-occurrences. In a similar spirit, our inclusion test is completed by searching the web for the missing (non-included) features on both sides. We call this web-based technique *mutual web-sampling*. The web results are further parsed to verify matching of the feature's syntactic relationship.

We denote the subset of *w*'s features that are missing for *v* as *M(w, v)* (and equivalently *M(v, w)*). Since web sampling is time consuming we randomly sample a subset of *k* features (*k=20* in our experiments), denoted as *M(v,w,k)*.

Mutual Web-sampling Procedure:

For each pair *(w, v)* and their *k*-subsets *M(w, v, k)* and *M(v, w, k)* execute:

1. Syntactic Filtering of "Bag-of-Words" Search:

Search the web for sentences including *v* and a feature *f* from *M(w, v, k)* as "bag of words", i. e. sentences where *w* and *f* appear in any distance and in either order. Then filter out the sentences that do not match the defined syntactic relation between *f* and *v* (based on parsing). Features that co-occur with *w* in the correct syntactic relation are removed from *M(w, v, k)*. Do the same search and filtering for *w* and features from *M(v, w, k)*.

2. Syntactic Filtering of "Exact String" Matching:

On the missing features on both sides (which are left in *M(w, v, k)* and *M(v, w, k)* after stage 1), apply "exact string" search of the web. For this, convert the tuple *(v, f)* to a string by adding prepositions and articles where needed. For example, for *(element, <project, pcomp_of, 1>)* generate the corresponding string "element of the project" and search the web for exact matches of the string. Then validate the syntactic relationship of *f* and *v* in the extracted sentences. Remove the found features from *M(w, v, k)* and *M(v, w, k)*, respectively.

3. Missing Features Validation:

Since some of the features may be too infrequent or corpus-biased, check whether the remaining missing features do co-occur on the web with their original target words (with which they did occur in the corpus data). Otherwise, they should not be considered as valid misses and are also removed from *M(w, v, k)* and *M(v, w, k)*.

Output: Inclusion in either direction holds if the corresponding set of missing features is now empty.

We also experimented with features consisting of words without syntactic relations. For example, exact string, or bag-of-words match. However, almost all the words (also non-entailing) were found with all the features of each other, even for semantically implausible combinations (e.g. a word and a feature appear next to each other but belong to different clauses of the sentence). Therefore we conclude that syntactic relation validation is very important, especially on the web, in order to avoid coincidental co-occurrences.

5 Empirical Results

To test the validity of the distributional inclusion hypotheses we performed an empirical analysis on a selected test sample using our automated testing procedure.

5.1 Data and setting

We experimented with a randomly picked test sample of about 200 noun pairs of 1,200 pairs produced by *RFF* (for details see Geffet and Dagan, 2004) under Lin's similarity scheme (Lin, 1998). The words were judged by the lexical entailment criterion (as described in Section 2). The original percentage of correct (52%) and incorrect (48%) entailments was preserved.

To estimate the degree of validity of the distributional inclusion hypotheses we decomposed each word pair of the sample *(w, v)* to two directional pairs ordered by potential entailment direction: *(w, v)* and *(v, w)*. The 400 resulting ordered pairs are used as a test set in Sections 5.2 and 5.3.

Features were computed from co-occurrences in a subset of the Reuters corpus of about 18 million words. For the web feature sampling the maximal number of web samples for each query (word - feature) was set to 3,000 sentences.

5.2 Automatic Testing the Validity of the Hypotheses at the Word Level

The test set of 400 ordered pairs was examined in terms of entailment (according to the manual judgment) and feature inclusion (according to the ITA algorithm), as shown in Table 2.

According to Hypothesis I we expect that a pair *(w, v)* that satisfies entailment will also preserve feature inclusion. On the other hand, by Hypothesis II if all the features of *w* are included by *v* then we expect that *w* entails *v*.

Entailment \ Inclusion	+	-
+	97	16
-	42	245

Table 2: Distribution of 400 entailing/non-entailing ordered pairs that hold/do not hold feature inclusion at the *word* level.

Entailment \ Inclusion	+	-
+	111	2
-	42	245

Table 4: Distribution of the entailing/non-entailing ordered pairs that hold/do not hold feature inclusion at the *sense* level.

spread – gap (mutually entail each other)
<weapon, pcomp_of>
The Committee was discussing the Programme of the "Big Eight," aimed against **spread of weapon** of mass destruction.

town – area ("town" entails "area")
<cooperation, pcomp_for>
This is a promising **area for cooperation** and exchange of experiences.

capital – town ("capital" entails "town")
<flow, nn>
Offshore financial centers affect cross-border **capital flow** in China.

Table 3: Examples of ambiguity of entailment-related words, where the disjoint features belong to a different sense of the word.

We observed that Hypothesis I is better attested by our data than the second hypothesis. Thus 86% (97 out of 113) of the entailing pairs fulfilled the inclusion condition. Hypothesis II holds for approximately 70% (97 of 139) of the pairs for which feature inclusion holds. In the next section we analyze the cases of violation of both hypotheses and find that the first hypothesis held to an almost perfect extent with respect to word senses.

It is also interesting to note that thanks to the web-sampling procedure over 90% of the non-included features in the corpus were found on the web, while most of the missing features (in the web) are indeed semantically implausible.

5.3 Manual Sense Level Testing of Hypotheses Validity

Since our data was not sense tagged, the automatic validation procedure could only test the hypotheses at the word level. In this section our goal is to analyze the findings of our empirical test at the word sense level as our hypotheses were defined for senses. Basically, two cases of hypotheses invalidity were detected:

Case 1: Entailments with non-included features (violation of Hypothesis I);

Case 2: Feature Inclusion for non-entailments (violation of Hypothesis II).

At the word level we observed 14% invalid pairs of the first case and 30% of the second case. However, our manual analysis shows, that over 90% of the first case pairs were due to a different sense of one of the entailing word, e.g. *capital - town* (*capital* as money) and *spread - gap* (*spread* as distribution) (Table 3). Note that ambiguity of the entailed word does not cause errors (like *town – area*, *area* as domain) (Table 3). Thus the first hypothesis holds at the sense level for over 98% of the cases (Table 4).

Two remaining invalid instances of the first case were due to the web sampling method limitations and syntactic parsing filtering mistakes, especially for some less characteristic and infrequent features captured by *RFF*. Thus, in virtually all the examples tested in our experiment Hypothesis I was valid.

We also explored the second case of invalid pairs: non-entailing words that pass the feature inclusion test. After sense based analysis their percentage was reduced slightly to 27.4%. Three possible reasons were discovered. First, there are words with features typical to the general meaning of the domain, which tend to be included by many other words of this domain, like *valley – town*. The features of *valley* ("eastern valley", "central valley", "attack in valley", "industry of the valley") are not discriminative enough to be distinguished from *town*, as they are all characteristic to any geographic location.

Method	*Precision*	*Recall*	*F1*
ITA-20	0.700	0.875	0.777
ITA-40	0.740	0.846	0.789
RFF-top-40	0.520	1.000	0.684
RFF-top-26	0.561	0.701	0.624

Table 5: Comparative results of using the filter, with 20 and 40 feature sampling, compared to *RFF* top-40 and *RFF* top-26 similarities. ITA-20 and ITA-40 denote the web-sampling method with 20 and random 40 features, respectively.

The second group consists of words that can be entailing, but only in a context-dependent (anaphoric) manner rather than ontologically. For example, *government* and *neighbour*, while *neighbour* is used in the meaning of *"neighbouring (country) government"*. Finally, sometimes one or both of the words are abstract and general enough and also highly ambiguous to appear with a wide range of features on the web, like *element (violence – element*, with all the tested features of *violence* included by *element)*.

To prevent occurrences of the second case more characteristic and discriminative features should be provided. For this purpose features extracted from the web, which are not domain-biased (like features from the corpus) and multi-word features may be helpful. Overall, though, there might be inherent cases that invalidate Hypothesis II.

6 Improving Lexical Entailment Prediction by ITA (Inclusion Testing Algorithm)

In this section we show that ITA can be practically used to improve the (non-directional) lexical entailment prediction task described in Section 2. Given the output of the distributional similarity method, we employ ITA at the word level to filter out non-entailing pairs. Word pairs that satisfy feature inclusion of all *k* features (at least in one direction) are claimed as entailing.

The same test sample of 200 word pairs mentioned in Section 5.1 was used in this experiment. The results were compared to *RFF* under Lin's similarity scheme (*RFF*-top-40 in Table 5).

Precision was significantly improved, filtering out 60% of the incorrect pairs. On the other hand, the relative recall (considering *RFF* recall as 100%) was only reduced by 13%, consequently leading to a better relative F1, when considering the *RFF*-top-40 output as 100% recall (Table 5).

Since our method removes about 35% of the original top-40 *RFF* output, it was interesting to compare our results to simply cutting off the 35% of the lowest ranked *RFF* words (top-26). The comparison to the baseline (*RFF*-top-26 in Table 5) showed that ITA filters the output much better than just cutting off the lowest ranking similarities.

We also tried a couple of variations on feature sampling for the web-based procedure. In one of our preliminary experiments we used the top-*k* *RFF* features instead of random selection. But we observed that top ranked *RFF* features are less discriminative than the random ones due to the nature of the *RFF* weighting strategy, which promotes features *shared* by many similar words. Then, we attempted doubling the sampling to 40 random features. As expected the recall was slightly decreased, while precision was increased by over 5%. In summary, the behavior of ITA sampling of *k=20* and *k=40* features is closely comparable (ITA-20 and ITA-40 in Table 5, respectively)[4].

7 Conclusions and Future Work

The main contributions of this paper were:

1. We defined two Distributional Inclusion Hypotheses that associate feature inclusion with lexical entailment at the word sense level. The Hypotheses were proposed as a refinement for Harris' Distributional hypothesis and as an extension to the classic distributional similarity scheme.
2. To estimate the empirical validity of the defined hypotheses we developed an automatic *inclusion testing algorithm* (ITA). The core of the algorithm is a web-based feature inclusion testing procedure, which helped significantly to compensate for data sparseness.
3. Then a thorough analysis of the data behavior with respect to the proposed hypotheses was conducted. The first hypothesis was almost fully attested by the data, particularly at the sense level, while the second hypothesis did not fully hold.
4. Motivated by the empirical analysis we proposed to employ ITA for the practical task of improving lexical entailment acquisition. The algorithm was applied as a filtering technique on the distributional similarity (*RFF*) output. We ob-

[4] The ITA-40 sampling fits the analysis from section 5.2 and 5.3 as well.

tained 17% increase of precision and succeeded to improve relative F1 by 15% over the baseline.

Although the results were encouraging our manual data analysis shows that we still have to handle word ambiguity. In particular, this is important in order to be able to learn the direction of entailment.

To achieve better precision we need to increase feature discriminativeness. To this end syntactic features may be extended to contain more than one word, and ways for automatic extraction of features from the web (rather than from a corpus) may be developed. Finally, further investigation of combining the distributional and the co-occurrence pattern-based approaches over the web is desired.

Acknowledgement

We are grateful to Shachar Mirkin for his help in implementing the web-based sampling procedure heavily employed in our experiments. We thank Idan Szpektor for providing the infrastructure system for web-based data extraction.

References

Chklovski, Timothy and Patrick Pantel. 2004. VERBOCEAN: Mining the Web for Fine-Grained Semantic Verb Relations. In Proc. of EMNLP-04. Barcelona, Spain.

Church, Kenneth W. and Hanks Patrick. 1990. Word association norms, mutual information, and Lexicography. Computational Linguistics, 16(1), pp. 22–29.

Dagan, Ido. 2000. Contextual Word Similarity, in Rob Dale, Hermann Moisl and Harold Somers (Eds.), Handbook of Natural Language Processing, Marcel Dekker Inc, 2000, Chapter 19, pp. 459-476.

Dagan, Ido, Oren Glickman and Bernardo Magnini. 2005. The PASCAL Recognizing Textual Entailment Challenge. In Proc. of the PASCAL Challenges Workshop for Recognizing Textual Entailment. Southampton, U.K.

Geffet, Maayan and Ido Dagan, 2004. Feature Vector Quality and Distributional Similarity. In Proc. of Coling-04. Geneva. Switzerland.

Grefenstette, Gregory. 1994. Exploration in Automatic Thesaurus Discovery. Kluwer Academic Publishers.

Harris, Zelig S. Mathematical structures of language. Wiley, 1968.

Hearst, Marti. 1992. Automatic acquisition of hyponyms from large text corpora. In Proc. of COLING-92. Nantes, France.

Keller, Frank, Maria Lapata, and Olga Ourioupina. 2002. Using the Web to Overcome Data Sparseness. In Jan Hajic and Yuji Matsumoto, eds., In Proc. of EMNLP-02. Philadelphia, PA.

Lee, Lillian. 1997. Similarity-Based Approaches to Natural Language Processing. Ph.D. thesis, Harvard University, Cambridge, MA.

Lin, Dekang. 1993. Principle-Based Parsing without Overgeneration. In Proc. of ACL-93. Columbus, Ohio.

.

Lin, Dekang. 1998. Automatic Retrieval and Clustering of Similar Words. In Proc. of COLING–ACL98, Montreal, Canada.

Moldovan, Dan, Badulescu, A., Tatu, M., Antohe, D., and Girju, R. 2004. Models for the semantic classification of noun phrases. In Proc. of HLT/NAACL-2004 Workshop on Computational Lexical Semantics. Boston.

Pado, Sebastian and Mirella Lapata. 2003. Constructing semantic space models from parsed corpora. In Proc. of ACL-03, Sapporo, Japan.

Pantel, Patrick and Dekang Lin. 2002. Discovering Word Senses from Text. In Proc. of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-02). Edmonton, Canada.

Pereira, Fernando, Tishby Naftali, and Lee Lillian. 1993. Distributional clustering of English words. In Proc. of ACL-93. Columbus, Ohio.

Ravichandran, Deepak and Eduard Hovy. 2002. Learning Surface Text Patterns for a Question Answering System. In Proc. of ACL-02. Philadelphia, PA.

Ruge, Gerda. 1992. Experiments on linguistically-based term associations. Information Processing & Management, 28(3), pp. 317–332.

Weeds, Julie and David Weir. 2003. A General Framework for Distributional Similarity. In Proc. of EMNLP-03. Sapporo, Japan.

Weeds, Julie, D. Weir, D. McCarthy. 2004. Characterizing Measures of Lexical Distributional Similarity. In Proc. of Coling-04. Geneva, Switzerland.