# Acquisition of Verb Entailment from Text

**Viktor Pekar**
Computational Linguistics Group
University of Wolverhampton
MB109 Stafford Street
Wolverhampton WV1 1SB, UK
v.pekar@wlv.ac.uk

## Abstract

The study addresses the problem of automatic acquisition of entailment relations between verbs. While this task has much in common with paraphrases acquisition which aims to discover semantic equivalence between verbs, the main challenge of entailment acquisition is to capture asymmetric, or directional, relations. Motivated by the intuition that it often underlies the local structure of coherent text, we develop a method that discovers verb entailment using evidence about discourse relations between clauses available in a parsed corpus. In comparison with earlier work, the proposed method covers a much wider range of verb entailment types and learns the mapping between verbs with highly varied argument structures.

## 1 Introduction

The entailment relations between verbs are a natural language counterpart of the commonsense knowledge that certain events and states give rise to other events and states. For example, there is an entailment relation between the verbs *buy* and *belong*, which reflects the commonsense notion that if someone has bought an object, this object belongs to that person.

A lexical resource encoding entailment can serve as a useful tool in many tasks where automatic inferencing over natural language text is required. In Question Answering, it has been used to establish that a certain sentence found in the corpus can serve as a suitable, albeit implicit answer to a query (Curtis et al., 2005), (Girju, 2003), (Moldovan and Rus, 2001). In Information Extraction, it can similarly help to recognize relations between named entities in cases when the entities in the text are linked by a linguistic construction that entails a known extraction pattern, but not by the pattern itself. A lexical entailment resource can contribute to information retrieval tasks via integration into a textual entailment system that aims to recognize entailment between two larger text fragments (Dagan et al., 2005).

Since entailment is known to systematically interact with the discourse organization of text (Hobbs, 1985), an entailment resource can be of interest to tasks that deal with structuring a set of individual facts into coherent text. In Natural Language Generation (Reiter and Dale, 2000) and Multi-Document Summarization (Barzilay et al., 2002) it can be used to order sentences coming from multiple, possibly unrelated sources to produce a coherent document. The knowledge is essential for compiling answers for procedural questions in a QA system, when sentences containing relevant information are spread across the corpus (Curtis et al., 2005).

The present paper is concerned with the problem of automatic acquisition of verb entailment from text. In the next section we set the background for the study by describing previous work. We then define the goal of the study and describe our method for verb entailment acquisition. After that we present results of its experimental evaluation. Finally, we draw conclusions and outline future work.

## 2 Previous Work

The task of verb entailment acquisition appears to have much in common with that of paraphrase acquisition (Lin and Pantel, 2001), (Pang et al., 2003), (Szpektor et al., 2004). In both tasks the goal is to discover pairs of related verbs and identify map-

*Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 49–56,
New York, June 2006. ©2006 Association for Computational Linguistics

pings between their argument structures. The important distinction is that while in a paraphrase the two verbs are semantically equivalent, entailment is a directional, or asymmetric, relation: one verb entails the other, but the converse does not hold. For example, the verbs *buy* and *purchase* paraphrase each other: either of them can substitute its counterpart in most contexts without altering their meaning. The verb *buy* entails *own* so that *buy* can be replaced with *own* without introducing any contradicting content into the original sentence. Replacing *own* with *buy*, however, does convey new meaning.

To account for the asymmetric character of entailment, a popular approach has been to use lexico-syntactic patterns indicative of entailment. In (Chklovski and Pantel, 2004) different types of semantic relations between verbs are discovered using surface patterns (like "*X-ed by Y-ing*" for enablement[1], which would match "*obtained by borrowing*", for example) and assessing the strength of asymmetric relations as mutual information between the two verbs. (Torisawa, 2003) collected pairs of coordinated verbs, i.e. matching patterns like "*X-ed and Y-ed*", and then estimated the probability of entailment using corpus counts. (Inui et al., 2003) used a similar approach exploiting causative expressions such as *because*, *though*, and *so*. (Girju, 2003) extracted causal relations between nouns like "*Earthquakes generate tsunami*" by first using lexico-syntactic patterns to collect relevant data and then using a decision tree classifier to learn the relations. Although these techniques have been shown to achieve high precision, their reliance on surface patterns limits their coverage in that they address only those relations that are regularly made explicit through concrete natural language expressions, and only within sentences.

The method for noun entailment acquisition by (Geffet and Dagan, 2005) is based on the idea of distributional inclusion, according to which one noun is entailed by the other if the set of occurrence contexts of the former subsumes that of the latter. However, this approach is likely to pick only a particular kind of verb entailment, that of troponymy (such as *march-walk*) and overlook pairs where there is little overlap in the occurrence patterns between the two verbs.

In tasks involving recognition of relations between entities such as Question Answering and Information Extraction, it is crucial to encode the mapping between the argument structures of two verbs. Pattern-matching often imposes restrictions on the syntactic configurations in which the verbs can appear in the corpus: the patterns employed by (Chklovski and Pantel, 2004) and (Torisawa, 2003) derive pairs of only those verbs that have identical argument structures, and often only those that involve a subject and a direct object. The method for discovery of inference rules by (Lin and Pantel, 2001) obtains pairs of verbs with highly varied argument structures, which also do not have to be identical for the two verbs. While the inference rules the method acquires seem to encompass pairs related by entailment, these pairs are not distinguished from paraphrases and the direction of relation in such pairs is not recognized.

To sum up, a major challenge in entailment acquisition is the need for more generic methods that would cover an unrestricted range of entailment types and learn the mapping between verbs with varied argument structures, eventually yielding resources suitable for robust large-scale applications.

## 3 Verb Entailment

Verb entailment relations have been traditionally attracting a lot of interest from lexical semantics research and their various typologies have been proposed (see, e.g., (Fellbaum, 1998)). In this study, with the view of potential practical applications, we adopt an operational definition of entailment. We define it to be a semantic relation between verbs where one verb, termed premise $P$, refers to event $E_p$ and at the same time implies event $E_q$, typically denoted by the other verb, termed consequence $Q$.

The goal of verb entailment acquisition is then to find two linguistic templates each consisting of a verb and slots for its syntactic arguments. In the pair, (1) the verbs are related in accordance with our definition of entailment above, (2) there is a mapping between the slots of the two templates and (3) the direction of entailment is indicated explic-

[1] In (Chklovski and Pantel, 2004) enablement is defined to be a relation where one event often, but not necessarily always, gives rise to the other event, which coincides with our definition of entailment (see Section 3).

itly. For example, in the template pair "*buy*(*obj:X*) ⇒ *belong*(*subj:X*)" the operator ⇒ specifies that the premise *buy* entails the consequence *belong*, and *X* indicates a mapping between the object of *buy* and the subject of *belong*, as in *The company bought shares. - The shares belong to the company.*

As opposed to logical entailment, we do not require that verb entailment holds in all conceivable contexts and view it as a relation that may be more plausible in some contexts than others. For each verb pair, we therefore wish to assign a score quantifying the likelihood of its satisfying entailment in some random context.

## 4 Approach

The key assumption behind our approach is that the ability of a verb to imply an event typically denoted by a different verb manifests itself in the regular co-occurrence of the two verbs inside locally coherent text. This assumption is not arbitrary: as discourse investigations show (Asher and Lascarides, 2003), (Hobbs, 1985), lexical entailment plays an important role in determining the local structure of discourse. We expect this co-occurrence regularity to be equally characteristic of any pair of verbs related by entailment, regardless of is type and the syntactic behavior of verbs.

The method consists of three major steps. First, it identifies pairs of clauses that are related in the local discourse. From related clauses, it then creates templates by extracting pairs of verbs along with relevant information as to their syntactic behavior. Third, the method scores each verb pair in terms of plausibility of entailment by measuring how strongly the premise signals the appearance of the consequence inside the text segment at hand. In the following sections, we describe these steps in more detail.

### 4.1 Identifying discourse-related clauses

We attempt to capture local discourse relatedness between clauses by a combination of several surface cues. In doing so, we do not build a full discourse representation of text, nor do we try to identify the type of particular rhetorical relations between sentences, but rather identify pairs of clauses that are likely to be discourse-related.

**Textual proximity**. We start by parsing the corpus with a dependency parser (we use Connexor's FDG (Tapanainen and Järvinen, 1997)), treating every verb with its dependent constituents as a clause. For two clauses to be discourse-related, we require that they appear close to each other in the text. Adjacency of sentences has been previously used to model local coherence (Lapata, 2003). To capture related clauses within larger text fragments, we experiment with windows of text of various sizes around a clause.

**Paragraph boundaries**. Since locally related sentences tend to be grouped into paragraphs, we further require that the two clauses appear within the same paragraph.

**Common event participant**. Entity-based theories of discourse (e.g., (Grosz et al., 1995)) claim that a coherent text segment tends to focus on a specific entity. This intuition has been formalized by (Barzilay and Lapata, 2005), who developed an entity-based statistical representation of local discourse and showed its usefulness for estimating coherence between sentences. We also impose this as a criterion for two clauses to be discourse-related: their arguments need to refer to the same participant, henceforth, **anchor**. We identify the anchor as the same noun lemma appearing as an argument to the verbs in both clauses, considering only subject, object, and prepositional object arguments. The anchor must not be a pronoun, since identical pronouns may refer to different entities and making use of such correspondences is likely to introduce noise.

### 4.2 Creating templates

Once relevant clauses have been identified, we create pairs of syntactic templates, each consisting of a verb and the label specifying the syntactic role the anchor occupies near the verb. For example, given a pair of clauses *Mary bought a house.* and *The house belongs to Mary.*, the method will extract two pairs of templates: {*buy*(*obj:X*), *belong*(*subj:X*)} and {*buy*(*subj:X*), *belong*(*to:X*).}

Before templates are constructed, we automatically convert complex sentence parses to simpler, but semantically equivalent ones so as to increase the amount of usable data and reduce noise:

- Passive constructions are turned into active ones: *X was bought by Y – Y bought X*;

- Phrases with coordinated nouns and verbs are decomposed: *X bought A and B – X bought A, X bought B; X bought and sold A – X bought A, X sold A.*
- Phrases with past and present participles are turned into predicate structures: *the group led by A – A leads the group; the group leading the market – the group leads the market.*

The output of this step is $V \in P \times Q$, a set of pairs of templates $\{p, q\}$, where $p \in P$ is the premise, consisting of the verb $v_p$ and $r_p$ – the syntactic relation between $v_p$ and the anchor, and $q \in Q$ is the consequence, consisting of the verb $v_q$ and $r_q$ – its syntactic relation to the anchor.

### 4.3 Measuring asymmetric association

To score the pairs for asymmetric association, we use a procedure similar to the method by (Resnik, 1993) for learning selectional preferences of verbs.

Each template in a pair is tried as both a premise and a consequence. We quantify the 'preference' of the premise $p$ for the consequence $q$ as the contribution of $q$ to the amount of information $p$ contains about its consequences seen in the data. First, we calculate Kullback-Leibler Divergence (Cover. and Thomas, 1991) between two probability distributions, $u$ – the prior distribution of all consequences in the data and $w$ – their posterior distribution given $p$, thus measuring the information $p$ contains about its consequences:

$$D_p(u||w) = \sum_n u(x) \log \frac{u(x)}{w(x)} \quad (1)$$

where $u(x) = P(q_x|p)$, $w(x) = P(q_x)$, and $x$ ranges over all consequences in the data. Then, the score for template $\{p, q\}$ expressing the association of $q$ with $p$ is calculated as the proportion of $q$'s contribution to $D_p(u||w)$:

$$Score(p, q) = P(q|p) \log \frac{P(q|p)}{P(p)} D_p(u||w)^{-1} \quad (2)$$

In each pair we compare the scores in both directions, taking the direction with the greater score to indicate the most likely premise and consequence and thus the direction of entailment.

## 5 Evaluation Design

### 5.1 Task

To evaluate the algorithm, we designed a recognition task similar to that of pseudo-word disambiguation (Schütze, 1992), (Dagan et al., 1999). The task was, given a certain premise, to select its correct consequence out of a pool with several artificially created incorrect alternatives.

The advantages of this evaluation technique are twofold. On the one hand, the task mimics many possible practical applications of the entailment resource, such as sentence ordering, where, given a sentence, it is necessary to identify among several alternatives another sentence that either entails or is entailed by the given sentence. On the other hand, in comparison with manual evaluation of the direct output of the system, it requires minimal human involvement and makes it possible to conduct large-scale experiments.

### 5.2 Data

The experimental material was created from the BLLIP corpus, a collection of texts from the Wall Street Journal (years 1987-89). We chose 15 transitive verbs with the greatest corpus frequency and used a pilot run of our method to extract 1000 highest-scoring template pairs involving these verbs as a premise. From them, we manually selected 129 template pairs that satisfied entailment.

For each of the 129 template pairs, four false consequences were created. This was done by randomly picking verbs with frequency comparable to that of the verb of the correct consequence. A list of parsed clauses from the BLLIP corpus was consulted to select the most typical syntactic configuration of each of the four false verbs. The resulting five template pairs, presented in a random order, constituted a test item. Figure 1 illustrates such a test item.

The entailment acquisition method was evaluated on entailment templates acquired from the British National Corpus. Even though the two corpora are quite different in style, we assume that the evaluation allows conclusions to be drawn as to the relative quality of performance of the methods under consideration.

```
1* buy(subj:X,obj:Y)⇒own(subj:X,obj:Y)
2  buy(subj:X,obj:Y)⇒approve(subj:X,obj:Y)
3  buy(subj:X,obj:Y)⇒reach(subj:X,obj:Y)
4  buy(subj:X,obj:Y)⇒decline(subj:X,obj:Y)
5  buy(subj:X,obj:Y)⇒compare(obj:X,with:Y)
```

Figure 1: An item from the test dataset. The template pair with the correct consequence is marked by an asterisk.

### 5.3 Recognition algorithm

During evaluation, we tested the ability of the method to select the correct consequence among the five alternatives. Our entailment acquisition method generates association scores for one-slot templates. In order to score the double-slot templates in the evaluation material, we used the following procedure.

Given a double-slot template, we divide it into two single-slot ones such that matching arguments of the two verbs along with the verbs themselves constitute a separate template. For example, "*buy* (*subj:X*, *obj:Y*) ⇒ *own* (*subj:X*, *obj:Y*)" will be decomposed into "*buy* (*subj:X*) ⇒ *own* (*subj:X*)" and "*buy* (*obj:Y*) ⇒ *own* (*obj:Y*)". The scores of these two templates are then looked up in the generated database and averaged. In each test item, the five alternatives are scored in this manner and the one with the highest score was chosen as containing the correct consequence.

The performance was measured in terms of accuracy, i.e. as the ratio of correct choices to the total number of test items. Ties, i.e. cases when the correct consequence was assigned the same score as one or more incorrect ones, contributed to the final accuracy measure proportionate to the number of tying alternatives.

This experimental design corresponds to a random baseline of 0.2, i.e. the expected accuracy when selecting a consequence template randomly out of 5 alternatives.

## 6 Results and Discussion

We now present the results of the evaluation of the method. In Section 6.1, we study its parameters and determine the best configuration. In Section 6.2, we compare its performance against that of human subjects as well as that of two state-of-the-art lexical resources: the verb entailment knowledge contained in WordNet2.0 and the inference rules from the DIRT database (Lin and Pantel, 2001).

### 6.1 Model parameters

We first examined the following parameters of the model: the window size, the use of paragraph boundaries, and the effect of the shared anchor on the quality of the model.

#### 6.1.1 Window size and paragraph boundaries

As was mentioned in Section 4.1, a free parameter in our model is a threshold on the distance between two clauses, that we take as an indicator that the clauses are discourse-related. To find an optimal threshold, we experimented with windows of 1, 2 ... 25 clauses around a given clause, taking clauses appearing within the window as potentially related to the given one. We also looked at the effect paragraph boundaries have on the identification of related clauses. Figure 2 shows two curves depicting the accuracy of the method as a function of the window size: the first one describes performance when paragraph boundaries are taken into account (PAR) and the second one when they are ignored (NO_PAR).

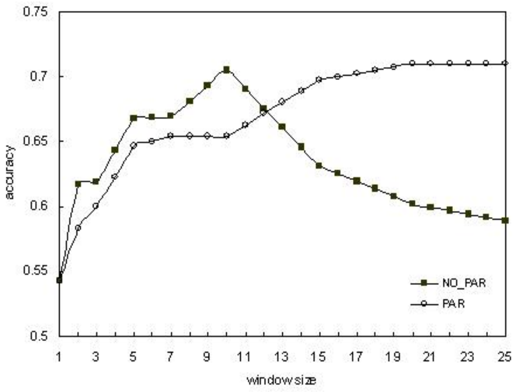

Figure 2: Accuracy of the algorithm as a function of window size, with and without paragraph boundaries used for delineating coherent text.

One can see that both curves rise fairly steeply up to window size of around 7, indicating that many entailment pairs are discovered when the two clauses appear close to each other. The rise is the steepest between windows of 1 and 3, suggesting that entailment relations are most often explicated in clauses appearing very close to each other.

PAR reaches its maximum at the window of 15, where it levels off. Considering that 88% of paragraphs in BNC contain 15 clauses or less, we take this as an indication that a segment of text where both a premise and its consequence are likely to be found indeed roughly corresponds to a paragraph. NO_PAR's maximum is at 10, then the accuracy starts to decrease, suggesting that evidence found deeper inside other paragraphs is misleading to our model.

NO_PAR performs consistently better than PAR until it reaches its peak, i.e. when the window size is less than 10. This seems to suggest that several initial and final clauses of adjacent paragraphs are also likely to contain information useful to the model.

We tested the difference between the maxima of PAR and NO_PAR using the sign test, the non-parametric equivalent of the paired t-test. The test did not reveal any significance in the difference between their accuracies (6-, 7+, 116 ties: p = 1.000).

#### 6.1.2 Common anchor

We further examined how the criterion of the common anchor influenced the quality of the model. We compared this model (ANCHOR) against the one that did not require that two clauses share an anchor (NO_ANCHOR), i.e. considering only co-occurrence of verbs concatenated with specific syntactic role labels. Additionally, we included into the experiment a model that looked at plain verbs co-occurring inside a context window (PLAIN). Figure 3 compares the performance of these three models (paragraph boundaries were taken into account in all of them).

Compared with ANCHOR, the other two models achieve considerably worse accuracy scores. The differences between the maximum of ANCHOR and those of the other models are significant according to the sign test (ANCHOR vs NO_ANCHOR: 44+, 8-, 77 ties: $p < 0.001$; ANCHOR vs PLAIN: 44+, 10-, 75 ties: $p < 0.001$). Their maxima are also reached sooner (at the window of 7) and thereafter their performance quickly degrades. This indicates that the common anchor criterion is very useful, especially for locating related clauses at larger distances in the text.

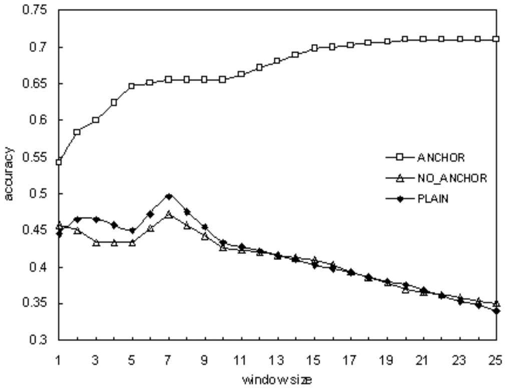

Figure 3: The effect of the common anchor on the accuracy of the method.

The accuracy scores for NO_ANCHOR and PLAIN are very similar across all the window size settings. It appears that the consistent co-occurrence of specific syntactic labels on two verbs gives no additional evidence about the verbs being related.

### 6.2 Human evaluation

Once the best parameter settings for the method were found, we compared its performance against human judges as well as the DIRT inference rules and the verb entailment encoded in the WordNet 2.0 database.

**Human judges**. To elicit human judgments on the evaluation data, we automatically converted the templates into a natural language form using a number of simple rules to arrange words in the correct grammatical order. In cases where an obligatory syntactic position near a verb was missing, we supplied the pronouns *someone* or *something* in that position. In each template pair, the premise was turned into a statement, and the consequence into a question. Figure 4 illustrates the result of converting the test item from the previous example (Figure 1) into the natural language form.

During the experiment, two judges were asked to mark those statement-question pairs in each test item, where, considering the statement, they could answer the question affirmatively. The judges' decisions coincided in 95 of 129 test items. The Kappa statistic is $\kappa$=0.725, which provides some indication about the upper bound of performance on this task.

```
X bought Y. After that:
1* Did X own Y?
2  Did X approve Y?
3  Did X reach Y?
4  Did X decline Y?
5  Did someone compare X with Y?
```

Figure 4: A test item from the test dataset. The correct consequence is marked by an asterisk.

**DIRT**. We also experimented with the inference rules contained in the DIRT database (Lin and Pantel, 2001). According to (Lin and Pantel, 2001), an inference rule is a relation between two verbs which are more loosely related than typical paraphrases, but nonetheless can be useful for performing inferences over natural language texts. We were interested to see how these inference rules perform on the entailment recognition task.

For each dependency tree path (a graph linking a verb with two slots for its arguments), DIRT contains a list of the most similar tree paths along with the similarity scores. To decide which is the most likely consequence in each test item, we looked up the DIRT database for the corresponding two dependency tree paths. The template pair with the greatest similarity was output as the correct answer.

**WordNet**. WordNet 2.0 contains manually encoded entailment relations between verb synsets, which are labeled as “cause”, “troponymy”, or “entailment”. To identify the template pair satisfying entailment in a test item, we checked whether the two verbs in each pair are linked in WordNet in terms of one of these three labels. Because WordNet does not encode the information as to the relative plausibility of relations, all template pairs where verbs were linked in WordNet, were output as correct answers.

Figure 5 describes the accuracy scores achieved by our entailment acquisition algorithm, the two human judges, DIRT and WordNet. For comparison purposes, the random baseline is also shown.

Our algorithm outperformed WordNet by 0.38 and DIRT by 0.15. The improvement is significant vs. WordNet (73+, 27-, 29 ties: p<0.001) as well as vs. DIRT (37+, 20-, 72 ties: p=0.034).

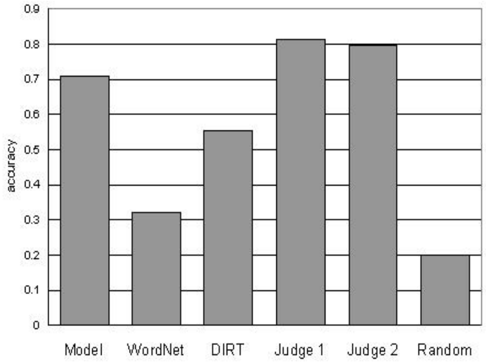

Figure 5: A comparison of performance of the proposed algorithm, WordNet, DIRT, two human judges, and a random baseline.

We examined whether the improvement on DIRT was due to the fact that DIRT had less extensive coverage, encoding only verb pairs with similarity above a certain threshold. We re-computed the accuracy scores for the two methods, ignoring cases where DIRT did not make any decision, i.e. where the database contained none of the five verb pairs of the test item. On the resulting 102 items, our method was again at an advantage, 0.735 vs. 0.647, but the significance of the difference could not be established (21+, 12-, 69 ties: p=0.164).

The difference in the performance between our algorithm and the human judges is quite large (0.103 vs. Judge 1 and 0.088 vs Judge 2), but significance to the 0.05 level could not be found (vs. Judge 1: 17-, 29+, 83 ties: p=0.105; vs. Judge 2: 15-, 27+, ties 87: p=0.09).

## 7 Conclusion

In this paper we proposed a novel method for automatic discovery of verb entailment relations from text, a problem that is of potential benefit for many NLP applications. The central assumption behind the method is that verb entailment relations manifest themselves in the regular co-occurrence of two verbs inside locally coherent text. Our evaluation has shown that this assumption provides a promising approach for discovery of verb entailment. The method achieves good performance, demonstrating a closer approximation to the human performance than inference rules, constructed on the basis of distributional similarity between paths in parse trees.

A promising direction along which this work can be extended is the augmentation of the current algorithm with techniques for coreference resolution. Coreference, nominal and pronominal, is an important aspect of the linguistic realization of local discourse structure, which our model did not take into account. As the experimental evaluation suggests, many verbs related by entailment occur close to one another in the text. It is very likely that many common event participants appearing in such proximity are referred to by coreferential expressions, and therefore noticeable improvement can be expected from applying coreference resolution to the corpus prior to learning entailment patterns from it.

## Acknowledgements

We are grateful to Nikiforos Karamanis and Mirella Lapata as well as three anonymous reviewers for valuable comments and suggestions. We thank Patrick Pantel and Dekang Lin for making the DIRT database available for this study.

## References

N. Asher and A. Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

R. Barzilay and M. Lapata. 2005. Modeling local coherence: an entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148.

R. Barzilay, N. Elhadad, and K. McKeown. 2002. Inferring strategies for sentence ordering in multidocument summarization. *JAIR*.

T. Chklovski and P. Pantel. 2004. VERBOCEAN: Mining the web for fine-grained semantic verb relations. In *In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*.

T.M. Cover. and J.A. Thomas. 1991. *Elements of Information Theory*. Wiley-Interscience.

J. Curtis, G. Matthews, and D. Baxter. 2005. On the effective use of cyc in a question answering system. In *Proceedings the IJCAI'05 Workshop on Knowledge and Reasoning for Answering Questions*.

I. Dagan, L. Lee, and F. Pereira. 1999. Similarity-based models of cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69.

I. Dagan, O. Glickman, and B. Magnini. 2005. The pascal recognising textual entailment challenge. In *PASCAL Challenges Workshop on Recognising Textual Entailment*.

C. Fellbaum, 1998. *WordNet: An Electronic Lexical Database*, chapter Semantic network of English verbs. MIT Press.

M. Geffet and I. Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 107–114.

R. Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL'03 Workshop on "Multilingual Summarization and Question Answering - Machine Learning and Beyond"*.

B. Grosz, A. Joshi, and S.Weinstein. 1995. Centering : a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

J.R. Hobbs. 1985. On the coherence and structure of discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information.

T. Inui, K.Inui, and Y.Matsumoto. 2003. What kinds and amounts of causal knowledge can be acquired from text by using connective markers as clues? In *Proceedings of the 6th International Conference on Discovery Science*, pages 180–193.

M. Lapata. 2003. Probabilistic text structuring: experiments with sentence ordering. In *Proceedings of the 41rd Annual Meeting of the Association for Computational Linguistics (ACL'03)*, pages 545–552.

D. Lin and P. Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360.

D. Moldovan and V. Rus. 2001. Logic form transformation of WordNet and its applicability to question answering. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*.

B. Pang, K. Knight, and D. Marcu. 2003. Syntax-based alignment of multiple translations: extracting paraphrases and generating new sentences. In *Proceedings of HLT-NAACL'2003*.

E. Reiter and R. Dale. 2000. *Building Natural Language Generation Systems*. Cambidge University Press.

P. Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

H. Schütze. 1992. Context space. In *Fall Symposium on Probabilistic Approaches to Natural Language*, pages 113–120.

I. Szpektor, H. Tanev, I. Dagan, and B. Coppola. 2004. Scaling web-based acquisition of entailment relations. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP'04)*.

P. Tapanainen and T. Järvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64–71.

K. Torisawa, 2003. *Questions and Answers: Theoretical and Applied Perspectives*, chapter An unsupervised learning method for commonsensical inference rules on events. University of Utrecht.