

Shallow Semantics for Textual Entailment Determination

Alina Andreevskaia, Zhuoyan Li and Sabine Bergler

CLaC Laboratory

Concordia University, Montreal,

{andreev, zhuoy_li, bergler}@cse.concordia.ca

Abstract

This paper analyses the contribution of shallow syntactic matching and thesaurus based equivalence in determining semantic equivalence of a pair of sentences. The performance of this approach is evaluated on two data sets and compared to other systems, as well as to manual evaluation results. We conclude that shallow semantics can model equivalence and entailment for pairs of syntactically similar sentences but it is not sufficient for reliable recognition of these relations in more complex cases.

1 Introduction

Textual entailment determination is one of the rapidly evolving areas of research in NLP. In this paper we adopt the definition of textual entailment advanced at the PASCAL recognizing textual entailment (RTE) challenge (2005) that was organized to explore what can be achieved in this area with current state-of-the-art tools. Thus, we adopt a definition of textual entailment as a directional relationship between pairs of text expressions, denoted by T (the entailing text), and H (the entailed hypothesis). It is considered that T entails H if the meaning of H can be inferred from the meaning of T , as would typically be inferred by people (Dagan et al., 2005).

Determination of textual entailment between two text segments (sentences in our case, see Table 1) is a fundamentally complex task that does not have an adequate solution up to date. Establishing the degree to which shallow semantics can con-

tribute to textual entailment evaluation can be a useful baseline for further research in this area. In order to explore the potential of these techniques, we have implemented a knowledge-lean system that uses shallow semantics to evaluate semantic equivalence between two sentences. The system has participated in the PASCAL RTE challenge (Dagan et al., 2005). Under shallow semantics we understand a combination of basic syntactic matching between the partial predicate-argument structures with simple thesaurus-based semantic equivalence. In our system, this semantic equivalence is measured using WordNet relations. In this paper we compare the performance of the system on the PASCAL data to the results obtained on the Microsoft paraphrase corpus, and contrast the performance of our system with theoretically developed lower bounds (Vanderwende et al., 2005). Additional experiments conducted for this paper give rise to more in-depth analysis of the potential and shortcoming of the shallow semantics approach to analysis of textual entailment.

2 Knowledge-lean textual entailment system implementation

The implementation of our textual entailment system is based on systems that the CLaC Laboratory developed for text summarization. The environment is implemented in the GATE architecture (Cunningham et al., 2002) and provides tagging, NP chunking, and knowledge-poor fuzzy NP coreference resolution as described in (Bergler et al., 2003), (Bergler et al., 2004), (Witte and Bergler, 2003). The flexible GATE architecture allows for the creation of modular components that can be used in different combinations depending on the task. For the pur-

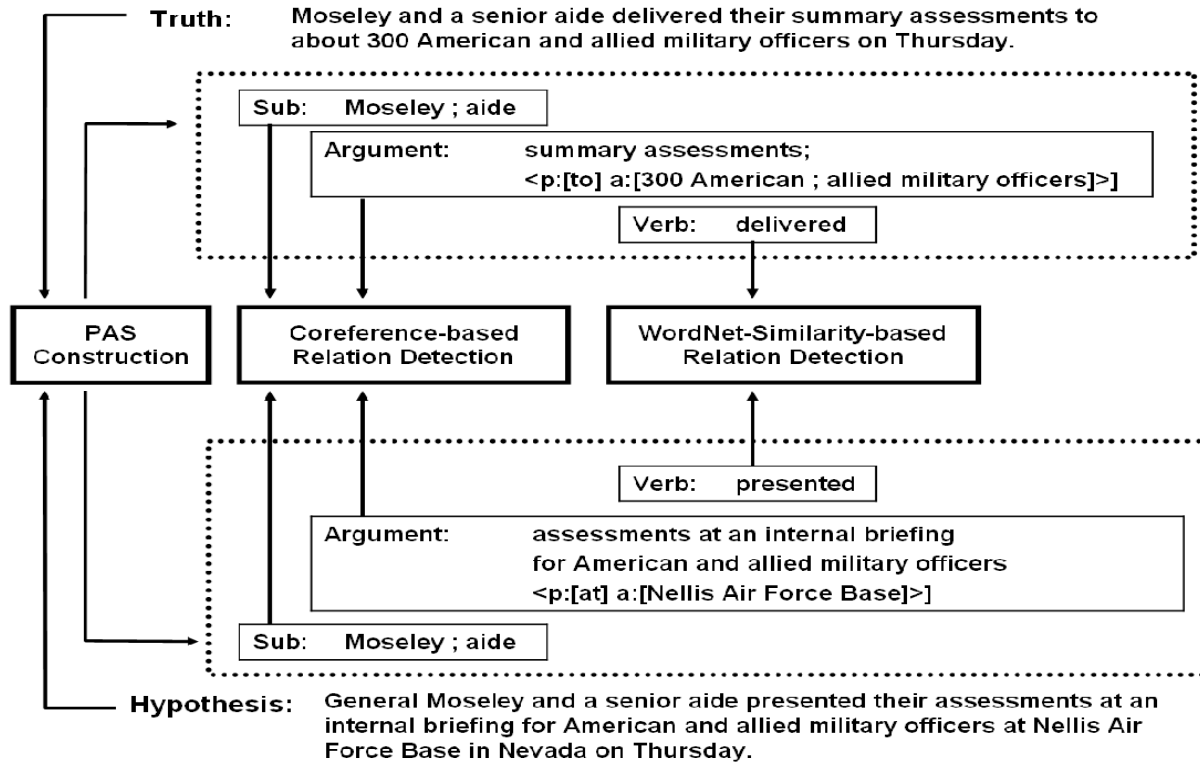


Figure 1: Example of a pair of paraphrases as processed by our system

poses of the textual entailment resolution, we extended the coreference system to incorporate verb groups, added full parsing, and included a few specialized heuristics for specific problems that were encountered in the PASCAL RTE challenge development set (Dagan et al., 2005). See (Andreevskaia et al., 2005) for details.

Our system uses shallow semantics to recognize equivalence between two sentences. Two main types of information were used to assess the relatedness between the two parts of the pair: partial Predicate Argument Structure (PAS), which is understood here as a representation of the meaning of the entire sentence, and lexical relatedness, which we measure using the WordNet distance (Figure 1). We use WordNet for lexical closeness only. No subcategorization information is used and thus our PASs are not necessarily complete.

Based on the results of the parsing, the system builds partial PAS for the two sentences that make up the pair (in the PASCAL RTE challenge notation: *T* or Text and *H* or Hypothesis). The constructed

PASs are compared constituent by constituent: subject to subject, verb to verb and object to object.¹ The comparison builds upon WordNet distance and coreference relations established by the fuzzy coreference module which includes pronoun resolution and string comparison components. Figure 1 illustrates these two techniques. The similarity between verbs was measured using WordNet distance that corresponds to the number of edges between two nodes (corresponding to synsets) in the tree. In Figure 1, *deliver* and *present* are members of the same synset in WordNet and thus the distance between them is 0. It is below the threshold (that can be set to any number ≥ 0) and therefore the words are considered semantically similar. Both subjects in the example are represented by the same strings and their

¹If the sentence contains more than one PAS all the predicate-argument structures are compared. In this case, comparison succeeds whenever a pair of similar PAS is detected. This approach is efficient and performs reasonably well but it may result in erroneous conclusions for cases where, for example, the main clause contains similar PASs while the subordinate clauses differ (as in speech reported from the same source).

	Text	Hypothesis
Syntax only MS corpus #1089874 and #1089925		
(1a)	PCCW’s chief operating officer, Mike Butcher, and Alex Arena, the chief financial officer, will report directly to Mr. So.	Current Chief Operating Officer Mike Butcher and Group Chief Financial Officer Alex Arena will report to So.
Syntax and lexical equivalence PASCAL test set pair #1361		
(1b)	A Filipino hostage in Iraq was released.	Filipino hostage was freed in Iraq.
World knowledge required PASCAL development set pair #98		
(1c)	Sharon warns Arafat could be targeted for assassination.	Prime minister targeted for assassination.

Table 1: Examples of pairs of equivalent sentences

equality in the two sentences is obvious. The equality of the direct objects is established based on the comparison of the heads of the corresponding noun phrases. The adjuncts (prepositional phrases in the boxes corresponding to arguments in Figure 1) were not considered in the comparison since they are not part of our predicate-argument structures.

Additional heuristics were implemented to deal with some typical patterns we encountered while developing the system. For example, we have implemented a special heuristic, termed *be-heuristic*, for sentences of type “*X is Y*” (Table 2). This heuristic first finds coreference chains in the first component of the pair (T) and establishes using a modified version of our multi-document coreference system whether any words in T corefer with words in H (based on string comparison). Then this information is used to prove whether the hypothesis that “*X is Y*” is true.

3 Data sets

The system was tested both on the PASCAL RTE challenge test set (further referred to as PASCAL

Text	Hypothesis
The centre-right European Peoples Party (EPP), the largest group in the European Parliament, has warned that it will reject the Taoiseach, Berni Ahern, if he is nominated as the next president of the European Commission.	Berni Ahern is the Taoiseach.

Table 2: Example of a pair processed by *be-heuristic* (PASCAL data; pair #336)

data) and on the test part of the Microsoft paraphrase corpus (MS corpus). The **Microsoft corpus** is a manually constructed set of text and hypothesis pairs of paraphrases. 66 % of these pairs are true equivalences. The corpus exhibits frequent structural and lexical similarity within the pairs. Some typical examples are given in Table 3.

	Text	Hypothesis
(2a)	Those conversations had not taken place as of Tuesday night, according to an Oracle spokeswoman.	Those talks have not taken place, according to an Oracle spokeswoman.
(2b)	The man accused of using fake grenades to commandeer a Cuban plane that landed in Key West in April was sentenced Friday to 20 years in prison.	A Cuban architect was sentenced to 20 years in prison Friday for using two fake grenades to hijack a passenger plane from Cuba to Florida in April.

Table 3: Examples of Microsoft corpus pairs

The **PASCAL data**, on the other hand, represents a set of manually matched sentences from different sources, grouped into seven categories: Comparable Documents (CD), Machine Translation (MT), Information Retrieval (IR) and Extraction (IE), Question Answering (QA), Reading Compre-

hension (RC) and Paraphrasing (PP). The categories differ in the way the data has been collected: for instance, in MT automatic translations were compared to standard human translations, in RC annotators manually created hypotheses corresponding to texts, while *T-H* pairs for CD task were selected from a cluster of comparable news articles. Each subset is made of equal number of true and false entailments. Certain pairs in the PASCAL data are very different in structure, as illustrated in Table 4.

	Text	Hypothesis
(3a)	Each hour spent in a car was associated with a 6 percent increase in the likelihood of obesity and each half-mile walked per day reduced those odds by nearly 5 percent, the researchers found.	The more driving you do means you're going to weigh more – the more walking means you're going to weigh less.
(3b)	Seven Egyptian human rights organisations issued a plea today, Monday, to Egyptian President Hosni Mubarak to hold accountable those responsible for the acts of torture that targeted residents of a village in Egypt's countryside while investigating two capital murders last August.	Particular seven organisations Egyptian Organisation for human rights today, Monday appealed to the Egyptian President Hosni Mubarak a cost-accounting the responsible for acts of torture which aimed villagers in upper Egypt during the investigation in the crimes killed in last August.

Table 4: Examples of PASCAL data pairs

Some pairs are difficult to evaluate even for humans. (Vanderwende et al., 2005) mentions 96% inter-annotator agreement for a sub-set of PASCAL data that can be evaluated based on syntax alone -

the part of the data that is one of the easiest for evaluation. These results are consistent with 91% inter-annotator agreement reported in (Bayer et al., 2005) for a sub-set of 70 pairs (10 pairs per task).

PASCAL data includes a small percentage (approximately 15 %) of entailments that are not equivalences (for example, Table 5).

Text	Hypothesis
Cedras, Biamby, and Francois also led the 1991 coup.	Cedras took part in the 1991 coup.

Table 5: Example of a true entailment that is not equivalence (PASCAL data; pair #1921)

The performance of our system on the MS corpus was considerably better than on the PASCAL data in terms of recall and precision but slightly worse for accuracy. This can be explained by the specifics of the composition of the two corpora: the PASCAL data is more heterogeneous and includes many different kinds of difficult cases (see Table 4). The MS corpus consists mostly of sentences that exhibit considerable similarity both in syntactic structure and lexical material, making our simple approach more successful (see Table 3).

4 Role of syntactic information

The system uses the similarity of partial syntactic structures of the sentences in the pair as the first cue of their equality. As mentioned in Section 2, we use full parse information. In fact, we use two parsers, the LINK parser (Sleator and Temperley, 1993) and the RASP parser (Briscoe and Carroll, 2002). We do not, however, match full syntactic structure in the pairs. Rather we approximate shallow semantics by constructing simplified shallow partial predicate-argument structures that cover only the verb, its subject and object (if there is one).

We use two parsers, because neither has a full coverage. One of the two parsers can be set as default, the second to be used only when the default parser doesn't produce a parse. Alternatively, both parsers can be given equal priority, and the system chooses for each sentence the parser that produces more PAs.

In order to evaluate the role of syntactic informa-

tion in the textual entailment determination a special run of the system without WordNet information was performed. For 51% of the PASCAL data and 49% of the MS corpus such basic PAS information was sufficient for the system to guess correctly (Table 6). The lower performance on the MS corpus is in part due to the corpus composition: it has only 33% “false” pairs and true negatives account for a little more than half of all correct guesses. For PASCAL data with its 50% of “false” pairs, the contribution of the true negatives is much greater - 83%. The system’s conservative strategy allows to pick up many true negatives, and misclassifying one of them has immediate impact on the system performance. 21%² of true equivalences in the MS corpus and 7% of true entailments in the PASCAL data were picked up by the system that used only syntactic information.

The composition of the two corpora in terms of syntactic similarity between the *T* and *H* parts has, in our opinion, even more significant impact on the results. As observed in Section 3, the pairs in the MS corpus are made of sentences similar in syntactic structure, while PASCAL data is considerably more heterogeneous. The only subset of PASCAL data that exhibits significant similarity between two components of a pair is the CD (Comparable Documents) task which has been processed by our system with much better results than other subsets of PASCAL data (Table 6). These numbers are closer to MS corpus results than to the overall system performance on PASCAL data, which is an additional argument in favor of this hypothesis.

corpus	A	CWS	P	R
MS test	0.49	0.48	0.79	0.32
PASCAL test	0.51	0.50	0.57	0.14
PASCAL-CD	0.61	0.57	0.92	0.15

Table 6: Accuracy (A), confidence-weighted score (CWS), precision (P) and recall (R) obtained using only syntactic information

Despite the use of very basic syntactic information, our approach still performs slightly better than approaches based on word overlap. For example, Perez and Alfonseca (2005) use BLEU to calculate

²Number of true positives over the total number of pairs.

how close two sentences are and reports 49.5% accuracy on the PASCAL data. At the same time, the two most successful systems submitted to PASCAL workshop - (Bayer et al., 2005) and (Glickman et al., 2005) - also don’t use any syntactic information but they employ corpus statistics and alignment techniques to capture additional information and both reach 0.586 accuracy.

Our results are consistent with manual estimates provided in (Vanderwende et al., 2005), who discuss a manual evaluation of the PASCAL data set. Linguistically trained human annotators judged whether a text-hypothesis pair of the data could be correctly identified as true or false based only on syntactic knowledge as given by current parsers, or whether it could be identified using a parser and a thesaurus. They report a baseline of 37% for purely syntax-driven approach, and 49% for the approach that uses a parser and a thesaurus. They claim that 10% of pairs in the PASCAL test data set can be judged as “true” based uniquely on syntactic information. Our number is lower for two reasons. First, we were considering only simplified predicate-argument structures and pronoun resolution, while Vanderwende et al. (2005) have a more broad definition of syntax, that included also a set of alternations³. Second, the estimates made by human annotators were based on idealized parses, while using real-life parsers leads to loss of information.⁴

Comparing the results of our system to the manual estimates for false entailments is more difficult. The annotators in (Vanderwende et al., 2005) were differentiating between cases where syntax could be a cause for evaluating the pair as false and where it wouldn’t. It turns out that distinguishing between the two is not straightforward and often causes disagreement between annotators. At the same time, we assign “false” to all pairs that are not considered true by the system and cannot tell, whether it is a sign that syntax is not sufficient for detection of entailment or whether mismatch of PASs was the reason for rejection. This difference in approaches

³We considered only passive-active alternation and partially incorporated promotion of appositive construction to main clause in the *be-heuristic*, which covers only a fraction of the possible transformations.

⁴Kouyelekov and Magnini (2005) report about 20% of sentences not getting correct parses.

would be the main reason for accuracy and CWS⁵ demonstrated by our system on the false entailments and on the entire set (51% compared to 37%).

Another factor that can potentially influence the performance of the system is the choice of the parser. We experimented with three different settings for the two parsers we use:

1. The one that produces most parses was chosen (called here *equal priority*).
2. Link parser is given higher priority meaning. It is used in all cases unless it fails to produce a parse.
3. RASP is given higher priority.

Equal priority resulted in marginally better performance on the PASCAL data, while giving higher priority to RASP slightly (0.5%) improved the precision and CWS on the MS corpus (Table 7). Overall, our experiments showed that the influence of the parser choice on the system results is negligible.

Corpus	Equal priority parsers			
	P	R	A	CWS
MS	0.79	0.34	0.50	0.49
PASCAL	0.59	0.13	0.52	0.52
Corpus	RASP/Link			
	P	R	A	CWS
MS	0.795	0.34	0.50	0.51
PASCAL	0.55	0.13	0.52	0.51
Corpus	Link/RASP			
	P	R	A	CWS
MS	0.79	0.34	0.50	0.51
PASCAL	0.58	0.12	0.52	0.51

Table 7: Parser priorities for WN distance=0

5 Role of lexical similarity

We used the WordNet (Fellbaum, 1998) to measure lexical similarity. We did that by computing the distance measured as the number of the edges between two nodes in the tree. This approach limits the application of lexical similarity to basic thesaural relations between words. Different thresholds were

⁵Confidence-Weighted Score (CWS) or Average Precision reflects the system’s ability to consistently assign higher confidence score to correct judgments (Dagan et al., 2005).

tested. The smaller values mean closer relationships, 0 corresponds to synonyms. Higher thresholds correspond to more permissive strategies and result in increase in recall. For PASCAL data (including the CD subset) this was achieved at the cost of decreased precision, while on the MS data precision didn’t suffer (Table 8), probably due to the fact that this corpus contains 66% true paraphrases, while PASCAL data has 50% true entailments.

Corpus	WN distance=0			
	P	R	A	CWS
MS	0.79	0.34	0.50	0.49
PASCAL	0.59	0.13	0.52	0.52
PASCAL-CD	0.92	0.31	0.64	0.64
Corpus	WN distance=1			
	P	R	A	CWS
MS	0.79	0.34	0.50	0.50
PASCAL	0.56	0.15	0.52	0.52
PASCAL-CD	0.92	0.31	0.64	0.64
Corpus	WN distance=2			
	P	R	A	CWS
MS	0.79	0.35	0.50	0.50
PASCAL	0.56	0.16	0.51	0.52
PASCAL-CD	0.89	0.32	0.65	0.64
Corpus	WN distance=3			
	P	R	A	CWS
MS	0.79	0.36	0.50	0.51
PASCAL	0.52	0.18	0.51	0.52
PASCAL-CD	0.83	0.32	0.63	0.64

Table 8: Results for varying WN thresholds

The comparison of the results on MS corpus, full PASCAL data and its CD subset demonstrates that the way the pairs are made up has considerable influence on the results. Higher precision and recall on MS corpus and especially on the CD task data are, in our opinion, due to the fact that these two data sets are composed of pairs that use lexical-level paraphrases that can be captured using WordNet relations.

Despite the increase in performance of the system when WordNet-based lexical similarity measure is introduced, the improvement was smaller than expected. Annotators in the study conducted by Vanderwende et al. (2005) estimated that syntax plus a basic thesaurus can handle 18% of true entailments

in PASCAL data (increase of 8% compared to making decisions based only on syntax). We were able to get only 2% increase compared to using only syntax and string matching. Similarly, only 4% increase was observed on the MS corpus. Overall, the gain in accuracy and CWS is also modest: the annotators estimated that 12% of correctly evaluated pairs can be added when basic semantic relations are taken into account, while we had only 2% increase in accuracy when WordNet-based similarity measures were considered. A possible explanation of such a difference is the inability of our system to handle synonyms that span over more than one word (for instance, *several people suffered injuries* vs. *several people were wounded*), which is necessary in many cases and which was considered as part of the general thesaurus by annotators in the (Vanderwende et al., 2005) study. Another reason for the poor performance of this approach is that it relies on the limited information contained in WordNet, while the two most successful systems submitted to PASCAL RTE challenge — (Bayer et al., 2005) and (Glickman et al., 2005) — used World Wide Web to extract word alignments (e.g., using WordNet synsets we can not capture the relationship between *outplay* and *defeat* that has been identified by Bayer et al. (2005), even though both words are in the dictionary).

6 System Limitations and Potential Improvements

The performance of the system that uses only shallow semantics is determined by several factors. The major stumbling block is the variations in the syntactic structure that are difficult to capture with partial syntactic information used by our system. Some of these variations represent recurring patterns that can be incorporated in a system as additional heuristics (as we have done for the "apposition $\rightarrow X$ is Y " pattern, that was implemented as *be-heuristic*). More such patterns have been reported in (Vanderwende et al., 2005). However, in most cases the difference in the syntactic structure can not be easily predicted (e.g., Table 9) and more sophisticated syntactic analysis is required for dealing with them.

The limited coverage of WordNet and the way its synsets have been constructed is another shortcoming of the simplified approach to establishing se-

Text	Hypothesis
By clicking here, you can return to the login page.	Click here to go back to the login page.

Table 9: Example of a pair made of syntactically different sentences (PASCAL data; pair #483)

semantic equivalence. For instance, multi-word paraphrases such as *fetus — unborn child* or *fire — send dismissal letters* can not be recognized using this data. Nevertheless, we believe that this information may be acquired from large corpora using statistical methods.

7 Conclusions

We have explored the contribution of shallow partial semantic analysis to the recognition of textual entailment and equivalence. We have validated results of our approach on two data sets - PASCAL RTE challenge data and Microsoft paraphrase corpus. We have also shown the results of shallow semantics approach to be in line with a theoretical study by Vanderwende et al. (2005) where human assessors stipulate whether entailment between two sentences can be established with syntax alone or with syntax plus a basic thesaurus.

The results of our experiments are comparable to the outcome of manual runs in (Vanderwende et al., 2005) when semantic similarity is not considered. Slightly lower results on true positives were mostly due to shortcomings of real-life parsers and to the limited number of alternations we consider. Adding WordNet-based semantic similarity measures did not increase the performance as much as it could have been expected based on results reported for human annotators. It demonstrates that shallow semantics is not sufficient for determining the equivalence and entailment in most cases and more sophisticated approaches are necessary.

The analysis of the system errors provided insights into promising directions for future research and system improvement. It can include more sophisticated syntactic analysis as well as acquiring more data on semantic equivalence at word level.

8 Acknowledgments

We are grateful to PASCAL RTE challenge organizers and Microsoft Research for sharing with us their corpora. We would like also to thank our anonymous reviewers for their valuable suggestions on the ways to improve this paper.

References

- Alina Andreevskaia, Zhuoy Li, and Sabine Bergler. 2005. Can Shallow Predicate-Argument Structures Determine Entailment? In *Proceedings of the First PASCAL Challenge Workshop for Recognizing Textual Entailment*, pages 45–49, <http://www.pascal-network.org/Challenges/RTE/>, 11-13 April. PASCAL.
- Samuel Bayer, John Burger, Lisa Ferro, John Henderson, and Alexander Yeh. 2005. MITREs Submissions to the EU Pascal RTE Challenge. In *Proceedings of the First PASCAL Challenge Workshop for Recognizing Textual Entailment*, pages 41–45, <http://www.pascal-network.org/Challenges/RTE/>, 11-13 April. PASCAL.
- Sabine Bergler, René Witte, Michelle Khalife, Zhuoyan Li, and Frank Rudzicz. 2003. Using Knowledge-poor Coreference Resolution for Text Summarization. In *Workshop on Text Summarization*, Document Understanding Conference (DUC), Edmonton, Canada, May 31–June 1. NIST.
- Sabine Bergler, René Witte, Michelle Khalife, Zhuoyan Li, Yunyu Chen, Monia Doandes, and Alina Andreevskaia. 2004. Multi-ERSS and ERSS 2004. In *Workshop on Text Summarization*, Document Understanding Conference (DUC), Boston, MA, May 6–7. NIST.
- E. Briscoe and J. Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1499–1504, Las Palmas, Canary Islands, May 2002.
- H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.
- Ido Dagan, Bernardo Magini, and Oren Glickman. 2005. The PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the First PASCAL Challenge Workshop for Recognizing Textual Entailment*, pages 1–9, <http://www.pascal-network.org/Challenges/RTE/>.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical database*. MIT Press.
- Oren Glickman, Ido Dagan, and Moshe Koppel. 2005. Web Based Probabilistic Textual Entailment. In *Proceedings of the First PASCAL Challenge Workshop for Recognizing Textual Entailment*, pages 33–37, <http://www.pascal-network.org/Challenges/RTE/>, 11-13 April. PASCAL.
- D. D. Sleator and D. Temperley. 1993. Parsing English with a link grammar. In *Third International Workshop on Parsing Technologies*.
- Lucy Vanderwende, Deborah Coughlin, and Bill Dolan. 2005. What Syntax can Contribute in Entailment Task. In *Proceedings of the First PASCAL Challenge Workshop for Recognizing Textual Entailment*, pages 13–17, <http://www.pascal-network.org/Challenges/RTE/>.
- René Witte and Sabine Bergler. 2003. Fuzzy Coreference Resolution for Summarization. In *Proceedings of 2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization (ARQAS)*, pages 43–50, Venice, Italy, June 23–24. Università Ca' Foscari.