Semantic Methods for Textual Entailment: How Much World Knowledge is Enough?

Andrew Neel and Max Garzon

Department of Computer Science
209 Dunn Hall
Memphis, Tennessee 38152 3240
aneel@memphis.edu, mgarzon@memphis.edu

Abstract

The problem of recognizing textual entailment (RTE) has been recently addressed with some success using semantic models. That attempt to capture the complexity of world knowledge. (Neel et al., 2008) has shown that semantic graphs made of synsets and selected relationships between them enable fairly simple methods to provide very competitive performance for RTE. Here, we extend the original results and show that RTE with automated word sense disambiguation (WSD) performs better using an updated WordNet database which presumably has evolved to capture more world knowledge than was available for the original evaluation. We obtain better results on datasets provided by subsequent RTE Challenge of 2008 and 2009. We report on the performance of these methods overall and in the four basic areas of information retrieval (IR), information extraction (IE), question answering (QA), and multi document summarization (SUM). We conclude that WordNet is not rich enough to provide appropriate information to resolve entailment with this inclusion protocol.

Introduction

The task of recognizing textual entailment (RTE) as defined by (Dagan et al. 2005; Bar-Haim et al. 2006) is the task of determining whether the *meaning* of one *text* (the *hypothesis*) is entailed (or inferred) from another text (simply, the *text*) to humans. It differs from logical inferences because world knowledge is required to assess a hypothesis. While performing RTE is fairly easy for humans for most instances of a problem, it is conversely difficult for computers in most instances.

The significance of finding a quality solution is high. Automatic solutions would have substantial impact on computers systems' capabilities on key textual entailment recognition tasks. Consider, for example, the task of automatic tutoring (Graesser et al. 2000; Graesser, Hu, and McNamara 2005), where a student provides answers to open ended questions asked by a tutoring system in natural language. Here, student answers must be evaluated against a number of known quality answers. Similarly, consider asking a computer to automatically summarize textual knowledge of several documents (called the multi-document summarization problem). Here, it may be desirable to remove sentences from the text which could be inferred by another sentence. A third example is information retrieval. Here, the goal is to find documents which are semantically similar in response to a query. Thus, the task for RTE is to evaluate the semantic closeness or relatedness of two documents and only provide a match when text documents do entail the query.

One related challenge to RTE is *Word Sense Disambiguation* (WSD). WSD literally applies knowledge of the world events to match words and phrases to meanings, herein called *synsets*. Similar to RTE, WSD is performed easily by humans but very poorly by computers. Thus, WSD also requires a protocol for disambiguating words and phrases into synsets *automatically, i.e.,* without human intervention or assistance.

Humans implicitly disambiguate words by matching the word *in context* to *meanings* and *experiences* stored in their memory. With humans, the context and experience serve as the world knowledge. Consider the following entailment instance: the text "*John Smith received a belt.*" entails the hypothesis *"John Smith received a strip of leather to fasten around his waist."* In this example, "belt" may have the meaning of "a strip of leather to fasten around the waste", "a strip of leather with machine-gun ammo attached" or "a strong punch". The human may remember the full set of *potential meanings* but experience will quickly identify "a strip of leather to fasten around the waste" as the specific and proper meaning. Resolving word/phrases to a list of synsets (i.e., a concept or meaning) is relatively easy. However, no automated solution has captured human experience sufficiently well to choose the appropriate meaning. Therefore, the crux of this issue is finding a representation of human world knowledge and experience in a model that will perform on computers the same function with comparable success, to humans.

In (Neel et Al, 2008), a protocol which used an automatic WSD solution was introduced to solve the RTE. The solution was shown to have good potential while remaining simple enough to implement on modern digital computers. The world knowledge provided by human experience was captured by WordNet, arguably capturing in a digital database a reasonable chunk of human experience (Kaplan and Shubert, 2001; Clark et Al., 2005). With Wordnet, the fundamental construct is not a word but an abstract semantic concept. Each concept, called *synset* (set of synonyms words), may be expressed by different words, and, conversely, the same word may represent different synsets. As the name implies, the concepts of WordNet are interconnected to provide a network of relationships between concepts. In (Neel et al., 2008), semantic models of world knowledge as exemplified by WordNet were used to show that semantic graphs made of synsets and selected relationships between them enable fairly simple methods that provide very competitive performance to WSD in the context of RTE problem instances. This solution to WSD then proved to be useful in performing automatic recognition of textual entailment in the four basic areas of information retrieval (IR), information extraction (IE), question answering (QA), and multi-document summarization (SUM). These datasets were provided by the then-current 2006 RTE challenge. The results were very competitive at the time with benchmark results as provided by the 2006 RTE challenge.

In this paper, we address a related problem, namely, how significant is the quality of world knowledge in solutions to RTE and WSD. It is clear that they have significant impact, but it is difficult to assess it precisely when both datasets and solutions are varying simultaneously, as it commonly the case when comparing solutions in the literature. For that purpose, the same methods used in (Neel et al., 2008) are applied to the more recent RTE data sets using more recent versions of the structure used to code for world knowledge which is presumably better. The conclusion in (Neel et al., 2008) suggested that results would improve automatically by simply improving the WordNet database. This work also addresses the chief criticism of the first study (Neel et al., 2008) where it was suggested that the datasets selected were not sufficiently diverse to evaluate the protocol properly. The implication was that by using only one dataset the results may have overly influenced by characteristics of that dataset. This paper shows that this is not so, i.e., that the original result holds as well, if not better, with datasets across the same four domains from two subsequent RTE Challenges Furthermore, the results also provide some quantification of the impact of world knowledge structures on RTE and WSD, including a discussion of how to quantify the quality of world knowledge ontologies.

. The paper is organized as follows: The next section provides the necessary background on the RTE Challenge and Word Sense Disambiguation. The following section recaps the protocols for WSD and RTE. Next, the original experiment of (Neel et al., 2008) was performed with an updated version of WordNet to assess its contribution. The following section evaluates our solution to RTE against two subsequent releases of the RTE Challenge datasets. Finally, we examine the contributions of WordNet and the inclusion protocol to solving WSD. In the conclusion section, the new results on WSD and RTE are used to attempt to quantify more precisely the impact of a given improvement on world knowledge structures.

Background and Related Information

RTE Challenge

A review of the RTE challenge is given in (Neel et al., 2008), but we summarize the basic experimental set up here in order to make this paper self-contained. In 2005, a first challenge was put forth (Dagan et al. 2005) to researchers to find a method that resolves or approximately resolves entailment. In order to be able to make objective comparisons between different solutions, a standard test data was published and has since been updated annually (Bar-Haim et al., 2006; Giampiccolo et al., 2007; Giampiccolo et al., 2008). Each challenge released datasets of 800 tuples across four domains or ontological categories. Each tuple consists of a “text” paragraph (T), at least one additional paragraph called “the hypothesis” (H), and the judgment about whether T entails H. The domains for each dataset include Information Extraction (IE), Information Retrieval (IR), Question Answer (QA), and Summarization (SUM). In (Neel et al., 2008), the authors selected a subset of 80 out of 800 tuples from the 2007 datasets (10 positive and 10 negative out of 100 tuples for each of the four domains) as a training set. From the training set, the authors determined an optimal threshold above which entailment could be assumed.

The TAC conference series (http://www.nist.gov/tac/) has held the challenge for two years and provided datasets with two important additional contributions. First, the length of the tuples is provided. The intuition is that some solutions may be better suited for shorter problems than longer problems. Second, decision on entailment was shifted from a binary decision (yes/no answer) to a three-way decision (yes, no, or “undetermined” answer). For this paper, the “undetermined” answers are removed from the 2008 dataset in order to make proper comparisons with the 2006 and 2007 datasets. Thus, the protocols discussed herein only identify entailment when it is clearly present but otherwise assumes that no entailment is present. In this paper, the accuracy of the protocol is evaluated against submissions to each challenge. Again, the results are further analyzed within the four separate domains described above and for its implications on the quality of the world knowledge structures used.

RTE using Automatic Word Disambiguation

A natural question about entailment is to quantify precisely the benefit that word sense disambiguation (WSD) has to entailment problems. This approach essentially ignores negation, sentence structure and resolves entailment solely on the contribution of WSD. (Neel et al., 2008) addressed this question by presenting a simple *inclusion* procedure to determine if a hypothesis (H) is entailed (or can be inferred) from a text (T) assuming word disambiguation has already been resolved. This protocol assumes that meanings of terms have been assessed *a priori*, presumably by humans. The algorithm determines the percentage of the overlap of synsets in the hypothesis with the synsets in the text. Table 2 (Neel et al., 2008) demonstrates the advantages of this approach. Key terms from two short paragraphs are disambiguated by human assessment using synsets provided by WordNet. In this example, *gunman* and *help* convey the same meaning as *shooters* and *aid*, respectively. However, lexical term matching will not identify the terms as a match. Pure lexical comparison would only match four of the six words in the hypothesis. By using the inclusion protocol for WSD, all terms were matched despite very different words. For conciseness, the term *inclusion protocol* will hereafter be used to refer to this protocol to evaluate entailment by automatic WSD using WordNet.

The *inclusion protocol* constructs a bi-partite graph for each tuple with one part corresponding to the text and the other to the hypothesis. The vertices are *synsets, not* words. Each part has an independent set of vertices where each represents one synset. The semantic relationships that would relate one word to another are represented by edges. Edges only connect vertices across regions. Both the semantic relationship and synsets associated with the text and hypothesis are determined by human or external assessment. Entailment is determined by how connected the synsets in part H are to the synsets in part T. The more H-vertices included in (connected to) T; the more likely it is that T entails H. Entailment is assumed to be false until enough of H connects to T. A threshold for how many connections are required to declare entailment present was optimized experimentally. This threshold was optimized with a training dataset, mentioned above, so that the solution shows the performance when word disambiguation was performed by human evaluation of entailment using this procedure.

	IE	IR	QA	SUM	Overall
Human WSD	45%	70%	67%	80%	72.90%
WordNet-2007	47%	65%	56%	70%	59.10%
WordNet-2009	47.5%	65.5%	57.5%	74.5%	61.25%
Improvement over 2007	2	1	5	9	17

Table 1: Accuracy of the inclusion protocol with automatic WSD using WordNet (Neel et al., 2008) increases with the quality of the world knowledge database. With the exception of IE, accuracy was significantly better than simply guessing. The overall accuracy of the 2007 evaluation outperformed 28 of 40 submissions to the 2006 RTE Challenge. The same procedure outperformed 35 of 40 using WordNet-2009.

In (Neel et al., 2008), human or external judgment was replaced with automated WSD procedures where the WordNet edition of 2007 (WordNet-2007) was provided the set of synsets. The automated process assumed that two words expressed the same meaning if either of them shared *any* word-meanings. Thus, the quality of the result was directly tied to the quality of the WordNet database.

Evaluating the Impact of WordNet

The inclusion procedure is a simple but clear demonstration that the human ability for assessing entailment depends on their capability to disambiguate words. This inclusion algorithm further shows that entailment could be assessed *automatically* by substituting the human assessment of WSD described in the above sections by a similar automatic WSD procedure.

The ideal for an automated WSD would be an evaluation of entailment that was very close to human WSD. In (Neel et al., 2008), the performance fell short by 13.8% overall. Interestingly, the performance of Information Extraction actually improved 2% over human WSD. The categories of Information Retrieval (IR), Question Answering (QA), and Summarization (SUM) were 5%, 11%, and 10% short respectively. This simple automated WSD would have enabled the inclusion protocol to best 28 of 40 submissions to the RTE challenge of 2006.

In the analysis of (Neel et al., 2008), it was reported that the shortcoming of performance for this automated WSD protocol can be traced back to the quality of WordNet, roughly described as the "quantity" of world knowledge. This conclusion is now be evaluated by substitution of WordNet-2007 with the November 2009 release of WordNet (hereafter called WordNet-2009) for further study.

Table 1 (row-1) (Human Evaluation) shows the performance of the protocol if the WSD is performed by humans. This data was data was collected as part of the evaluation presented in (Neel et al., 2008).

Table 1 (row-2) shows the performance of entailment using this protocol. Three of the four categories of data determined entailment better than just random guessing, which is expected to be correct about 50% of the time. Information retrieval (IR) and summarization (SUM) were more than 15% above the performance of simple guessing. Overall scores show a near 10% improvement over guessing. Only one case (IE) performed worse than guessing (by more than 5%). When compared to the same protocol using human disambiguation (Table 1 row-1), the scores for Information Extraction (IE) do *improve by* 2%. The remaining three categories performed worse by 5-10%

when compared with the protocol that assumes disambiguation. Overall scores are about 10% worse.

Next, WordNet-2007 was replaced with a November-2009 edition, which is essentially the same database; but with 2+ years of additional world knowledge. The same protocols for evaluation of entailment and for word-sense disambiguation were e used. Only the source of world knowledge was exchanged. The first two sources included are human judgment of word-sense meaning and the 2007 edition of WordNet. The results for these two evaluations were first presented in (Neel et al., 2008). The third source is based on the more recent 2009 edition of WordNet.

The performance of the inclusion protocol improved across every category. Table 1 (row-3) (WordNet-2009) shows the accuracy of the inclusion protocol with automatic WSD using WordNet-2009. IE improved by evaluating 1 more tuple as an entailment. IR, QA, and SUM improved by 1, 5, and 9 (resp.). Overall, WordNet-2009 enabled correct entailment in 17 more cases. As before, this result was compared with those of the RTE2 challenge submissions. With WordNet-2009, we found that this simple inclusion protocol with our automatic WSD protocol now out-performed 35 (7 more) out of 40 submissions.

Further Evaluation of Inclusion Protocol

Here, the inclusion protocol is expanded from RTE2 in 2006 (Bar-Haim et Al. 2006) to the subsequent RTE Challenge datasets of RTE3 in 2007 (Giampiccolo et Al., 2007) and RTE4 in 2008 (Giampiccolo et Al., 2008). Table 2 compares the accuracy of the inclusion protocol using automatic WSD for RTE2, RTE3, and RTE4 (columns). The accuracy is reported for each domain of Information Extraction (IE), Information Retrieval (IR), question answering (QA), summarization (SUM), and overall performance (rows). (The results of RTE2 are also presented in Table 1 and are simply repeated here for completeness.) The performance here would have outscored 35 of 40 submissions to the RTE2 challenge.

The protocol was performed without any modification on the RTE3 dataset (meaning that no modification was made at all to the WSD protocol or inclusion protocol for RTE.) Here, the protocol performed better with RTE3 (column-2 of Table 2) in every domain except SUM. Further, the overall performance RTE3 was essentially identical to that of RTE2 (column-1 of Table 2). The only notable changes were in the domains of QA and SUM where the protocol's accuracy increased from 57.5% to 68.5% for QA and decreased from 74.5% to 57.5% for SUM. This result would have scored 27 of 45 submissions to the RTE3 challenge.

The protocol was again performed without modification on the RTE4 dataset. Here, the protocol performed worse with RTE4 (column-3 of Table 2) in every domain except IE. The overall performance decreased to slightly better than guessing. The full set of scores of submission to the RTE4 challenge was not available for comparison. Consequently, it was not possible to place our score in context with other submissions. However, the top eight scores were available for submissions making only a two-way decision (meaning they submission only assessed whether entailment was present or not). In comparison to the eight reported scores, the protocol performed 8% worse than the lowest of the top eight scores. (Giampiccolo et Al., 2008) reported the average per domain as 52% for IE, 60% for SUM, 61% for IR, and 52% for QA. Here, our protocol beat the average for IE and tied with the average for QA.

	RTE2	**RTE3**	**RTE4**	***RTE4 (modified)***
IE	47.5%	52.0%	54.9%	*63.1%*
IR	65.5%	66.5%	48.2%	*54.4%*
QA	57.5%	68.5%	51.5%	*69.2%*
SUM	74.5%	57.5%	52.3%	*63.1%*
Overall	61.3%	61.1%	51.7%	*61.7%*

Table 2: The accuracy of the inclusion protocol using automatic WSD for RTE2, RTE3, and RTE4 (columns) is compared. The accuracy is reported for each and overall performance (rows). The performance of RTE2 and RTE3 are identical while the performance of RTE4 is not as good. When thresholds for assessing entailment are relaxed to account for missing information in WordNet, the result improves substantially.

An analysis of this result of RTE4 was performed focusing on WSD. It was found that the RTE2 and RTE3 datasets contained approximately 6000 words. Of these words, WordNet matched all but 900 for RTE2 and all but 1000 for RTE3. The dataset of RTE4 contained no less than 8000 words (2000 additional words) than either RTE2 or RTE3 datasets. Of the 8000 words, 1500 were not found in WordNet. Therefore it is reasonable to conclude that WordNet was not able to provide enough information to resolve entailment with the inclusion protocol.

Next, the inclusion protocol itself was examined. By examining each tuple closely and comparing it with the result of the entailment assessment from the inclusion protocol, it was determined that the number of false negatives substantially outnumbered that of false-positives (the protocol returned a result of no-entailment when entailment was really present.) This result further emphasized that automatic WSD did not provide enough information to assess entailment.

To test this result, thresholds in the inclusion protocol were decreased by 10%. Consequently, the protocol would return entailment if the hypothesis contained 10% fewer matches than before. Column-5 of Table 2 shows the result of this assessment. The performance improved in each domain to either tie or better the average. The overall performance was sufficient to move into 7th place in the two-way challenge.

Impact of Quality of World Knowledge

Here, we evaluate the efficiency of the inclusion protocol to perform WSD and draw some conclusions on the impact of quality of world-knowledge models on WSD. The inclusion protocol relies on WordNet to provide world knowledge. This paper has evaluated an automated method for WSD and used that automated WSD method to assess entailment using the inclusion protocol. Having now examined the inclusion protocol over several datasets and compared the performance of the protocol with two versions of WordNet, we can now evaluate the inclusion protocol with automated WSD in light of the new results. The product in this evaluation provides evidence of WordNet's shortcomings in capturing/providing world knowledge and the inclusion protocols ability to efficiently use the information available through WordNet. Thus, two questions need to be answered: (1) "Does WordNet contain enough world knowledge to enable the Inclusion protocol to perform automated WSD in order to successfully assess entailment?" and (2) "Is the world knowledge provided by WordNet used to its maximum potential?". Neither the inclusion protocol nor automated WSD were performed on the RTE3 and RTE4 for lack of access to the previous version of WordNet. Therefore, no data was available to provide comparisons. These data points are marked with an 'x' in Tables 3.

Table 3A shows the number of words extracted from each RTE challenge. The number of words extracted for RTE2 was the same in 2007 as it was in 2009 since only a new version of WordNet was substituted for this experiment. In 2007, 4,788 words (or 80.7% of the words) matched (or hit) anything in WordNet. By 2009, WordNet had matured to match 5,015 words (or 84.5%). Further analysis showed that all 4,788 hits from 2007 were found to be included in 5,015 hits in 2009 with no modification. Thus, the only additional contribution of WordNet between 2007 and 2008 was the 227 new words relevant to RTE2. This observation suggests that incomplete entries are not made into WordNet and that WordNet's content is relatively stable once it is entered. The words extracted from RTE3 and RTE4 challenge datasets hit 83.6% and 82.1%. Even though the number of words increased year over year, the percentage of words with any word-synset relationship declines year to year. Overall, the average word-synset matches in WordNet are between 80-85%. Assuming that entries are found in WordNet are stable and complete, this result would indicate that WordNet does capture a large portion of world knowledge.

Table 3A also shows that the total number of words and the total number of words without any match in WordNet increased from RTE2 to RTE3 and RTE4. RTE3 performed essentially the same as RTE2 having 260 more words. RTE4 had 1967 more words than RTE3 and 2227 more words than RTE2 and resulted in the inclusion protocols performance being far worse with RTE4 than RTE2 or RTE3. As discussed above, after the obvious step of optimizing the thresholds used by the inclusion protocol, the performance of RTE4 returned to match that of RTE2 and RTE3. Consequently, the inclusion protocol itself appears to perform comparatively well even when world knowledge is less than perfect. In fact, the result emphasizes the sensitivity of the inclusion protocol to the availability of world knowledge.

Table 3B shows the count of relationships between words and synsets. The totals for synsets are shown in Table 5A. The 227 additional words produced an additional 1,397 relationships which resulted in an additional 441 synsets. The automated WSD of the inclusion protocol works by assuming two words convey the same meaning if they share a relationship to common synsets (hence the inclusion of the synsets of one text in

(A) Items Provided from WordNet

	RTE2	RTE3	RTE4
Words	5,935	6,195	8,162
WordNet-2007 Hits	4,788 (80.7%)	x	x
WordNet-2009 Hits	5,015 (84.5%)	5,181 (83.6%)	6,697 (82.1%)

(B) Word-Synset Relationships

	RTE2	RTE3	RTE4
Wordnet-2007	28,601	x	X
Wordnet-2009	29,999	30,302	38,032

(C) Synsets

	RTE2	RTE3	RTE4
Wordnet-2007	16,807	x	X
Wordnet-2009	17,248	17,392	21,007

(D) Avg Number of Synsets Per Word

	RTE2	RTE3	RTE4
Wordnet-2007	2.83	x	x
Wordnet-2009	2.91	2.81	2.57

Table 3: (A) provides a count of the words extracted from each of the RTE2, RTE3, and RTE4 datasets (respectively). WordNet provides some world knowledge in 80-85% of the time. The ratio of words to matches in WordNet declines with the newer datasets. (B) provides a count of the word-synset relationships extracted from WordNet for each dataset. The number of relationships increases substantially with later versions of WordNet. (C) provides a count of the number of synsets and, thus, provides a relative measure of world knowledge captured. (D) shows the average number of synsets per word. Despite its decline, the performance of the inclusion algorithm remains steady with the same threshold naturally declines.

another) (Neel et al., 2008). Since the 4,788 words that hit in WordNet for RTE2 had 29,999 word-synset relationships but only 17,248 synsets, we can conclude that a fairly large percentage of the words in the RTE challenge conveyed the same meaning as other words used while the words used to convey those meanings were substantially different. In fact, there are almost twice as many word-synset relationships as synsets in the portion of WordNet. This result clearly shows that the automated WSD and the inclusion protocol are efficiently using the information retrieved from WordNet.

Table 3D shows the average number of synsets per word. In 2007, there were 5,935 words and 28,601 synsets for an average of 2.83 words per synset. It only improved to 2.91 in 2009. RTE3 and RTE4 had the values of 2.81 and 2.57 respectively in 2009. This value provides a rough measure of the knowledge extracted from each word. As the value increases (as it did with RTE2 from 2007 to 2009), the inclusion protocol assesses entailment far more accurately and efficiently. Further, as the value decreases (as it does from RTE2 to RTE3/RTE4 datasets), the inclusion protocol becomes far less efficient, but can be .

The addition of 227 word-synset relationships translated into a 2.15% increase in inclusion protocols evaluation efficiency of entailment on RTE2. The increase results is manifest as more information becomes available to the automated WSD protocol used by the inclusion protocol. However, even with a word-synset hit-rate of nearly 85%, the inclusion protocol was short 11.65% (Table 3) from WSD by human assessment by. This result seems to suggest that the inclusion protocol will perform much better as WordNet improves but will fall short of human assessment of WSD even if 100% of World knowledge is contained in word-synset relationships of WordNet.

The original experiment of (Neel et al., 2008) tested hypernymy and trophonymy relationships as a possible avenue for enhancing the efficiency of the automated WSD protocol. In the first experiment, 3,946 of the 4,788 words of RTE2 with word-synset relationships had hypernymy or trophonymy relationships. The total number of hypernymy or trophonymy relationships totaled 25,147 and the total hypernymy or trophonymy synsets totaled 8,353. In 2009, 4,466 of the 5,015 words with word-synset relationships had hypernymy or trophonymy relationships. Here, the total number of relationships increased to 2,109,460 and the total hypernymy or trophonymy synsets totaled 20,306. Thus, WordNet has grown substantially in the complexity of its relationships. Though (Neel et al., 2008) showed that hypernymy or trophonymy relationships did not help and actually hurt performance in a few cases, substantial growth in WordNet's complexity in this area may warrant a second examination of these types of relationships.

Conclusions and Future Work

In this paper we have extended prior results of a new (inclusion) protocol for automatics solutions of the RTE challenge and Word Sense Disambiguation (WSD) to more recent RTE data sets and presumably better world knowledge encoding in the form of WordNet's ontologies. The evidence presented here shows the inclusion protocol for assessing entailment that works well when substantial world knowledge is available for word sense disambiguation. The accuracy of the RTE protocol degrades as the quality or availability of world knowledge degrades. This results confirms the significance of world knowledge encoding in RTE and WSD. The RTE protocol remains very competitive with minor adjustments, despite including only word-meaning (synsets) and despite a significant decrease in world knowledge per word provided in more recent version of WordNet.,

It may be possible to improve performance by incorporating additional semantic relationships captured in WordNet, or better yet, better models of world knowledge. Furthermore. as pointed out by (Neel et al., 2008), incorporating other type of world knowledge in the form of language structure (such as negation, sentence structure, and other conventional solutions), may further improve the effectiveness of the inclusion protocols for WSD and RTE.

References

1. Bar Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B. & Szpektor, I. (2006). The Second PASCAL RTE Challenge. *In Proceedings of the 2nd PASCAL Challenge on RTE*.
2. Clark, P., Harrison, P., Jenkins, T., Thompson, J., Wojcik, R. (2005). Acquiring and Using World Knowledge using a Restricted Subset of English. *In Proceedings of the eighteenth International Florida AI Research Symposium.* FLAIRS, AAAI 2005. Clearwater Beach, Florida, USA. 506 511.
3. Dagan, I., Glickman. O., & Magnini, B. The PASCAL RTE Challenge. *In Proceedings of the 1st PASCAL Challenge Workshop on RTE*.
4. Giampiccolo, D., Magnini, B., Dagan, I., and Dolan, B. (2007). The 3rd PASCAL Recognizing Textual Entailment Challenge. *In Proc. of the Third PASCAL Challenges Workshop on RTE*.
5. Giampiccolo, D., Dang, H.T., Magnini, B., Dagan, I., Cabrio, E., and Dolan, B. (2008). The Fourth PASCAL Recognizing Textual Entailment Challenge. *In Proceedings of the Third PASCAL Challenges Workshop on RTE*.
6. Graesser, A., Wiemer Hastings, P., Wiemer Hastings, K., Harter, D.,Person, N., and the Tutoring Research Group. (2000). Using Latent Semantic Analysis to evaluate the contributions of students in AutoTutor. Interactive Learning Environments, 8, 149 169.
7. Kaplan, A.N., and Schubert, L.K. 2001. Measuring and improving the quality of world knowledge extracted from WordNet. Tech. Rep. 751 14627 0226, Dept. of Computer Science, Univ. of Rochester, Rochester, NY, May.
8. Neel, A., Garzon, M.H., & Rus, V. (2008). A Semantic Method for Textual Entailment. *In Proceedings of the Twenty First International Florida AI Research Symposium*. FLAIRS, AAAI 2008. Coconut Grove, Florida, USA. 171 176.
9. Raina, R., Ng, A., & Manning, C.D. (2005). Robust textual inference via learning and abductive reasoning. *Proc. of the 20th National Conf. on Artificial Intelligence.* AAAI Press.