

Recognising Entailment within Discourse

Shachar Mirkin[§], Jonathan Berant[†], Ido Dagan[§], Eyal Shnarch[§]

[§] Computer Science Department, Bar-Ilan University

[†] The Blavatnik School of Computer Science, Tel-Aviv University

Abstract

Texts are commonly interpreted based on the entire discourse in which they are situated. Discourse processing has been shown useful for inference-based application; yet, most systems for textual entailment – a generic paradigm for applied inference – have only addressed discourse considerations via off-the-shelf coreference resolvers. In this paper we explore various discourse aspects in entailment inference, suggest initial solutions for them and investigate their impact on entailment performance. Our experiments suggest that discourse provides useful information, which significantly improves entailment inference, and should be better addressed by future entailment systems.

1 Introduction

This paper investigates the problem of recognising textual entailment within discourse. Textual Entailment (TE) is a generic framework for applied semantic inference (Dagan et al., 2009). Under TE, the relationship between a *text* (T) and a textual assertion (*hypothesis*, H) is defined such that *T entails H* if humans reading T would infer that H is most likely true (Dagan et al., 2006).

TE has been successfully applied to a variety of natural language processing applications, including information extraction (Romano et al., 2006) and question answering (Harabagiu and Hickl, 2006). Yet, most entailment systems have thus far paid little attention to discourse aspects of inference. In part, this is the result of the unavailability of adept tools for handling the kind of discourse processing required for inference. In addition in the main TE benchmarks, the Recognising Textual Entailment (RTE) challenges, discourse

played little role. This state of affairs has started to change with the recent introduction of the RTE Pilot “Search” task (Bentivogli et al., 2009b), in which assessed texts are situated within complete documents. In this setting, texts need to be interpreted based on their entire discourse (Bentivogli et al., 2009a), hence attending to discourse issues becomes essential. Consider the following example from the task’s dataset:

(T) *The seven men on board were said to have as little as 24 hours of air.*

For the interpretation of T, e.g. the identity and whereabouts of the seven men, one must consider T’s discourse. The preceding sentence T’, for instance, provides useful information to that aim:

(T’) *The Russian navy worked desperately to save a small military submarine.*

This example demonstrates a common situation in texts, and is also applicable to the RTE Search task’s setting. Still, little was done by the task’s participants to consider discourse, and sentences were mostly processed independently.

Analyzing the Search task’s development set, we identified several key discourse aspects that affect entailment in a discourse-dependent setting. First, we observed that the coverage of available coreference resolution tools is considerably limited. To partly address this problem, we extend the set of coreference relations to phrase pairs with a certain degree of lexical overlap, as long as no semantic incompatibility is found between them. Second, many bridging relations (Clark, 1975) are realized in the form of “global information” perceived as known for entire documents. As bridging falls completely out of the scope of available resolvers, we address this phenomenon by identifying and weighting prominent document terms and allowing their incorporation in inference even

when they are not explicitly mentioned in a sentence. Finally, we observed a coherence-related discourse phenomenon, namely inter-relations between entailing sentences in the discourse, such as the tendency of entailing sentences to be adjacent to one another. To that end, we apply a two-phase classification scheme, where a second-phase meta-classifier is applied, extracting discourse and document-level features based on the classification of each sentence on its own.

Our results show that, even when simple solutions are employed, the reliance on discourse-based information is helpful and achieves a significant improvement of results. We analyze the contribution of each component and suggest some future work to better attend to discourse in entailment systems. To our knowledge, this is the most extensive effort thus far to empirically explore the effect of discourse on entailment systems.

2 Background

Discourse plays a key role in text understanding applications such as question answering or information extraction. Yet, such applications typically only handle a narrow aspect of discourse, addressing coreference by term substitution (Dali et al., 2009; Li et al., 2009). The limited coverage and scope of existing tools for coreference resolution and the unavailability of tools for addressing other discourse aspects also contribute to this situation. For instance, VP anaphora and bridging relations are usually not handled at all by such resolvers. A similar situation is seen in the TE research field.

The prominent benchmark for entailment systems evaluation is the series of RTE challenges. The main task in these challenges has traditionally been to determine, given a text-hypothesis pair (T,H), whether T entails H. Discourse played no role in the first two RTE challenges as T's were constructed of short simplified texts. In RTE-3 (Giampiccolo et al., 2007), where some paragraph-long texts were included, inter-sentential relations became relevant for correct inference. Yet the texts in the task were manually modified to ensure they are self-contained. Consequently, little effort was invested by the challenges' participants to address discourse issues beyond the standard substitution of coreferring

nominal phrases, using publicly available tools such as JavaRap (Qiu et al., 2004) or OpenNLP¹, e.g. (Bar-Haim et al., 2008).

A major step in the RTE challenges towards a more practical setting of text processing applications occurred with the introduction of the Search task in the Fifth RTE challenge (RTE-5). In this task entailing sentences are situated within documents and depend on other sentences for their correct interpretation. Thus, discourse becomes a substantial factor impacting inference. Surprisingly, discourse hardly received any treatment in this task beyond the standard use of coreference resolution (Castillo, 2009; Litkowski, 2009), and an attempt to address globally-known information by removing from H words that appear in document headlines (Clark and Harrison, 2009).

3 The RTE Search Task

The RTE-5 Search task was derived from the TAC Summarization task². The dataset consists of several corpora, each comprised of news articles concerning a specific *topic*, such as the impact of global warming on the Arctic or the London terrorist attacks in 2005. *Hypotheses* were manually generated based on Summary Content Units (Nenkova et al., 2007), clause-long statements taken from manual summaries of the corpora. *Texts* are unmodified sentences in the articles. Given a topic and a hypothesis, entailment systems are required to identify all sentences in the topic's corpus that entail the hypothesis.

Each sentence-hypothesis pair in both the development and test sets was annotated, judging whether the sentence entails the hypothesis. Out of 20,104 annotations in the development set, only 810 were judged as positive. This small ratio (4%) of positive examples, in comparison to 50% in traditional RTE tasks, better corresponds to the natural distribution of entailing texts in a corpus, thus better simulates practical settings.

The task may seem as a variant of information retrieval (IR), as it requires finding specific texts in a corpus. Yet, it is fundamentally different from IR for two reasons. First, the target output is a set

¹<http://opennlp.sourceforge.net>

²<http://www.nist.gov/tac/2009/Summarization/>

of sentences, each evaluated independently, rather than a set of documents. Second, the decision criterion is *entailment* rather than *relevance*.

Despite the above, apparently, IR techniques provided hard-to-beat baselines for the RTE Search task (MacKinlay and Baldwin, 2009), outperforming every other system that relied on inference without IR-based pre-filtering. At the current state of performance of entailment systems, it seems that lexical coverage largely overshadows any other approach in this task. Still, most (6 out of 8) participants in the challenge applied their entailment systems to the entire dataset without a prior retrieval of candidate sentences. F₁ scores for such systems vary between 10% and 33%, in comparison to over 40% of the IR-based methods.

4 The Baseline RTE System

In this work we used BIUTEE, Bar-Ilan University Textual Entailment Engine (Bar-Haim et al., 2008; Bar-Haim et al., 2009), a state of the art RTE system, as a baseline and as a basis for our discourse-based enhancements. This section describes this system’s architecture; the methods by which it was augmented to address discourse are presented in Section 5.

To determine entailment, BIUTEE performs the following main steps:

Preprocessing First, all documents are parsed and processed with standard tools for named entity recognition (Finkel et al., 2005) and coreference resolution. For the latter purpose, we use OpenNLP and enable the substitution of corefering terms. This is the only way by which BIUTEE addresses discourse, representing the state of the art in entailment systems.

Entailment-based transformations Given a T-H pair (both represented as dependency parse trees), the system applies a sequence of knowledge-based entailment transformations over T, generating a set of texts which are entailed by it. The goal is to obtain consequent texts which are more similar to H. Based on preliminary results on the development set, in our experiments (Section 6) we use WordNet (Fellbaum, 1998) as the system’s only knowledge resource, using its synonymy, hyponymy and derivation relations.

Classification A supervised classifier, trained on the development set, is applied to determine entailment of each pair based on a set of syntactic and lexical syntactic features assessing the degree by which T and its consequents cover H.

5 Addressing Discourse

In the following subsections we describe the prominent discourse phenomena that affect inference, which we have identified in an analysis of the development set and addressed in our implementation. As mentioned, these phenomena are poorly addressed by available reference resolvers or fall completely out of their scope.

5.1 Augmented coreference set

A large number of coreference relations are comprised of terms which share lexical elements, (e.g. “*airliners’s first flight*” and “*Airbus A380’s first flight*”). Although common in coreference relations, standard resolvers miss many of these cases. For the purpose of identifying additional corefering terms, we consider two noun phrases in the same document as corefering if: (i) their heads are identical and (ii) no semantic incompatibility is found between their modifiers. The types of incompatibility we handle are: (a) mismatching numbers, (b) antonymy and (c) co-hyponymy (coordinate terms), as specified by WordNet. For example, two nodes of the noun *distance* would be considered incompatible if one is modified by *short* and the second by its antonym *long*. Similarly, two modifier co-hyponyms of *distance*, such as *walking* and *running* would also result such an incompatibility. Adding more incompatibility types (e.g. *first* vs. *second flight*) may further improve the precision of this method.

5.2 Global information

Key terms or prominent pieces of information that appear in the document, typically at the title or the first few sentences, are many times perceived as “globally” known throughout the document. For example, the geographic location of the document theme, mentioned at the beginning of the document, is assumed to be known from that point on, and will often not be mentioned explicitly in further sentences. This is a bridging phenomenon

that is typically not addressed by available discourse processing tools. To compensate for that, we identify key terms for each document based on *tf-idf* scores and consider them as global information for that document. For example, global terms for the topic discussing the ice melting in the Arctic, typically contain a location such as *Arctic* or *Antarctica* and terms referring to *ice*, like *permafrost* or *iceshelf*.

We use a variant of *tf-idf*, where term frequency is computed as follows: $tf(t_{i,j}) = n_{i,j} + \vec{\lambda}^\top \cdot \vec{f}_{i,j}$. Here, $n_{i,j}$ is the frequency of term i in document j ($t_{i,j}$), which is incremented by additional positive weights ($\vec{\lambda}$) for a set of features ($\vec{f}_{i,j}$) of the term. Based on our analysis, we defined the following features, which correlated mostly with global information: (i) does the term appear in the title? (ii) is it a proper name? (iii) is it a location? The weights for these features are set empirically.

The document’s top- n global terms are added to each of its sentences. As a result, a global term that occurs in the hypothesis is matched in each sentence of the document, regardless of whether the term explicitly appears in the sentence.

Considering the previous sentence Another method for addressing missing coreference and bridging relations is based on the assumption that adjacent sentences often refer to the same entities and events. Thus, when extracting classification features for a given sentence, in addition to the features extracted from the parse tree of the sentence itself, we extract the same set of features from the current and previous sentences together. Recall the example presented in Section 1. T is annotated as entailing the hypothesis “*The AS-28 mini-submarine was trapped underwater*”, but the word *submarine*, e.g., appears only in its preceding sentence T’. Thus, considering both sentences together when classifying T increases its coverage of the hypothesis. Indeed, a bridging reference relates *on board* in T with *submarine* in T’, justifying our assumption in this case.

5.3 Document-level classification

Beyond discourse references addressed above, further information concerning discourse and document structure is available in the Search setting

and may contribute to entailment classification. We observed that entailing sentences tend to come in bulks. This reflects a common coherence aspect, where the discussion of a specific topic is typically continuous rather than scattered across the entire document. This *locality* phenomenon may be useful for entailment classification since knowing that a sentence entails the hypothesis increases the probability that adjacent sentences entail the hypothesis as well.

To capture this phenomenon, we use a two-phase meta-classification scheme, in which a *meta-classifier* utilizes entailment classifications of the first classification phase to extract *meta-features* and determine the final classification decision. This scheme also provides a convenient way to combine scores from multiple classifiers used in the first classification phase. We refer to these as *base-classifiers*. This scheme and the meta-features we used are detailed hereunder.

Let us write (s, h) for a sentence-hypothesis pair. We denote the set of pairs in the development (training) set as \mathcal{D} and in the test set as \mathcal{T} . We split \mathcal{D} into two halves, \mathcal{D}_1 and \mathcal{D}_2 . We make use of n base-classifiers, C_1, \dots, C_n , among which C^* is a designated classifier with additional roles in the process, as described below. Classifiers may differ, for example, in their classification algorithm. An additional meta-classifier is denoted C_M . The classification scheme is shown as Algorithm 1.

Algorithm 1 Meta-classification

Training

- 1: Extract features for every (s, h) in \mathcal{D}
- 2: Train C_1, \dots, C_n on \mathcal{D}_1
- 3: Classify \mathcal{D}_2 , using C_1, \dots, C_n
- 4: Extract meta-features for \mathcal{D}_2 using the classification of C_1, \dots, C_n
- 5: Train C_M on \mathcal{D}_2

Classification

- 6: Extract features for every (s, h) in \mathcal{T}
 - 7: Classify \mathcal{T} using C_1, \dots, C_n
 - 8: Extract meta-features for \mathcal{T}
 - 9: Classify \mathcal{T} using C_M
-

At Step 1, features are extracted for every (s, h) pair in the training set, as in the baseline system.

In Steps 2 and 3 we split the training set into two halves (taking half of each topic), train n different classifiers on the first half and then classify the second half using each of the n classifiers. Given the classification scores of the n base-classifiers to the (s, h) pairs in the second half of the training set, \mathcal{D}_2 , we add in Step 4 the meta-features described in Section 5.3.1.

After adding the meta-features, we train (Step 5) a meta-classifier on this new set of features. Test sentences then go through the same process: features are extracted for them and they are classified by the already trained n classifiers (Steps 6 and 7), meta-features are extracted in Step 8, and a final classification decision is made by the meta-classifier in Step 9.

A retrieval step may precede the actual entailment classification, allowing the processing of fewer and potentially “better” candidates.

5.3.1 Meta-features

The following features are extracted in our meta-classification scheme:

Classification scores The classification score of each of the n base-classifiers.

Title entailment In many texts, and in news articles in particular, the title and the first few sentences often represent the entire document’s content. Thus, knowing whether these sentences entail the hypothesis may be an indicator to the general potential of the document to include entailing sentences. Two binary features are added according to the classification of C^* indicating whether the title entails the hypothesis and whether the first sentence entails it.

Second-closest entailment Considering the locality phenomenon described above, we add a feature assigning higher scores to sentences in the vicinity of an entailment environment. This feature is computed as the distance to the second-closest entailing sentence in the document (counting the sentence itself as well), according to the classification of C^* . Formally, let i be the index of the current sentence and \mathcal{J} be the set of indices of entailing sentences in the document according to C^* . For each $j \in \mathcal{J}$ we compute $d_{i,j} = |i-j|$, and choose the second smallest $d_{i,j}$ as d_i . The idea is

#	Ent?	Closest	d	2nd closest	d
1	NO	6	5	7	6
2	NO	6	4	7	5
3	NO	6	3	7	4
4	NO	6	2	7	3
5	NO	6	1	7	2
6	YES	7	1	7	1
7	YES	6 or 8	1	6 or 8	1
8	YES	7 or 9	1	7 or 9	1
9	YES	8	1	8	1
10	NO	8	1	8	2
11	NO	8	2	8	3

Figure 1: Comparison of the *closest* and *second-closest* schemes when applied to a bulk of entailing sentences (in white) situated within a non-entailing environment (in gray). Unlike the *closest* one, the *second-closest* scheme assigns larger distance values to non-entailing sentences located on the ‘edge’ of the bulk (5 and 10) than to entailing ones.

that if entailing sentences indeed always come in bulks, then $d_i = 1$ for all entailing sentences, but $d_i > 1$ for all non-entailing ones. Figure 1 illustrates such a case, comparing the *second-closest* distance with the distance to the *closest* entailing sentence. In the *closest* scheme we do not count the sentence as closest to itself since it would disregard the environment of the sentence altogether, eliminating the desired effect. We scale the distance and add the feature score: $-\log(d_i)$.

Smoothed entailment This feature addressed the locality phenomenon by smoothing the classification score of sentence i with the scores of adjacent sentences, weighted by their distance from the current sentence i . Let $s(i)$ be the score assigned by C^* to sentence i . We add the Smoothed Entailment feature score:

$$SE(i) = \frac{\sum_w (b^{|w|} \cdot s(i+w))}{\sum_w (b^{|w|})}$$

where $0 < b < 1$ is the decay parameter and w is an integer bounded between $-N$ and N , denoting the distance from sentence i .

1st sentence entailing title Bensley and Hickl (2008) showed that the first sentence in a news article typically entails the article’s title. We therefore assume that in each document, $s_1 \Rightarrow s_0$, where s_1 and s_0 are the document’s first sentence and title respectively. Hence, under entailment transitivity, if $s_0 \Rightarrow h$ then $s_1 \Rightarrow h$. The corresponding binary feature states whether the sentence being classified is the document’s first sentence *and* the title entails h according to C^* .

	P (%)	R (%)	F ₁ (%)
<i>BIU-BL</i>	14.53	55.25	23.00
<i>BIU-DISC</i>	20.82	57.25	30.53
<i>BIU-BL</i> ³	14.86	59.00	23.74
<i>BIU-DISC</i> _{no-loc}	22.35	57.12	32.13
All-yes baseline	4.6	100.0	8.9

Table 1: Micro-average results.

Note that the above locality-based features rely on high accuracy of the base classifier C^* . Otherwise, it will provide misleading information to the features computation. We analyze the effect of this accuracy in Section 6.

6 Results and Analysis

Using the RTE-5 Search data, we compare BIUTEE in its baseline configuration (cf. Section 4), denoted *BIU-BL*, with its discourse-aware enhancement (*BIU-DISC*) which uses all the components described in Section 5. To alleviate the strong IR effect described in Section 3, both systems are applied to the complete datasets (both training and test), without candidates pre-filtering.

BIU-DISC uses three base-classifiers ($n = 3$): SVM^{perf} (Joachims, 2006), and Naïve Bayes and Logistic Regression from the WEKA package (Witten and Frank, 2005). The first among these is set as our designated classifier C^* , which is used for the computation of the document-level features. SVM^{perf} is also used for the meta-classifier. For the smoothed entailment score (cf. Section 5.3), we used $b = 0.9$ and $N = 3$. Global information is added by enriching each sentence with the highest-ranking term in the document, according to *tf-idf* scores (cf. Section 5.2), where document frequencies were computed based on about half a million documents from the TIPSTER corpus (Harman, 1992). The set of weights $\vec{\lambda}$ equals $\{2, 1, 4\}$ for title terms, proper names and locations, respectively. All parameters were tuned based on a 10-fold cross-validation on the development set, optimizing the micro-averaged F_1 .

The results are presented in Table 1. As can be seen in the table, *BIU-DISC* outperforms *BIU-BL* in every measure, showing the impact of addressing discourse in this setting. To rule out the option that the improvement is simply due to the fact that we use three classifiers for *BIU-DISC* and a single one

	P (%)	R (%)	F ₁ (%)
By Topic			
<i>BIU-BL</i>	16.54	55.62	25.50
<i>BIU-DISC</i>	22.69	57.96	32.62
All-yes baseline	4.85	100.00	9.25
By Hypothesis			
<i>BIU-BL</i>	22.87	59.62	33.06
<i>BIU-DISC</i>	27.81	61.97	38.39
All-yes baseline	4.96	100.00	9.46

Table 2: Macro-average results.

for *BIU-BL*, we show (*BIU-BL*³) the results when the baseline system is applied in the same meta-classification configuration as *BIU-DISC*, with the same three classifiers. Apparently, without the discourse information this configuration’s contribution is limited.

As mentioned in Section 5.3, the benefit from the locality features rely directly on the performance of the base classifiers. Hence, considering the low precision scores obtained here, we applied *BIU-DISC* to the data in the meta-classification scheme, but with locality features removed. The results, shown as *BIU-DISC*_{no-loc} in the Table, indicate that indeed performance increases without these features. The last line of the table shows the results obtained by a naïve baseline where all test-set pairs are considered entailing.

For completeness, Table 2 shows the macro-averaged results, when averaged over the topics or over the hypotheses. Although we tuned our system to maximize micro-averaged F_1 , these figures comply with the ones shown in Table 1.

Analysis of locality As discussed in Section 5, determining whether a sentence entails a hypothesis should take into account whether adjacent sentences also entail the hypothesis. In the above experiment we were unable to show the contribution of our system’s component that attempts to capture this information; on the contrary, the results show it had a negative impact on performance.

Still, we claim that this information can be useful when used within a more accurate system. We try to validate this conjecture by understanding how performance of the locality features varies as the systems becomes more accurate. We do so via the following simulation.

When classifying a certain sentence, the classi-

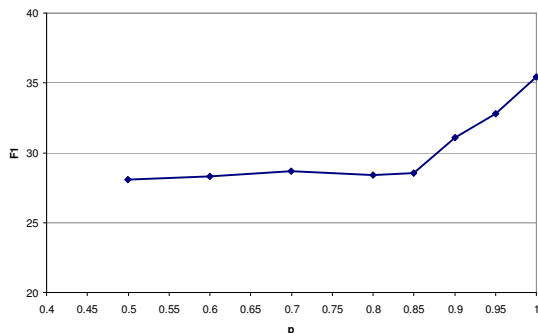


Figure 2: F_1 performance of *BIU-DISC* as a function of the accuracy in classifying adjacent sentences.

fications of its adjacent sentences are given by an *oracle classifier* that provides the correct answer with probability p . The system is applied using two locality features: the *1st sentence entailing title* feature and a close variant of the *smoothed entailment* feature, which calculates the weighted average of adjacent sentences, but disregards the score of the currently evaluated sentence.³ Thus we supply information about adjacent sentences and test whether overall performance increases with the accuracy of this information.

We performed this experiment for p in a range of [0.5-1.0]. Figure 2 shows the results of this simulation, based on the average F_1 of five runs for each p . Since performance, from a certain point, increases with the accuracy of the oracle classifier, we can conclude that indeed precise information about adjacent sentences improves performance on the current sentence, and that locality is a true phenomenon in the data. We note, however, that performance improves only when accuracy is very high, suggesting the currently limited practical potential of this information, at least in the way locality was represented in this work.

Ablation tests Table 3 presents the results of the ablation tests performed to evaluate the contribution of each component. Based on the result reported in Table 1 and the above discussion, the tests were performed relative to *BIU-DISC_{no-loc}*, the optimal configuration. As seen in the table, the removal of each component causes a drop in results. For global information we see a mi-

³The *second-closest entailment* feature was not used as it considers the oracle’s decision for the current sentence, while we wish to use information about adjacent sentences.

Component removed	F_1 (%)	ΔF_1 (%)
Previous sent. features	28.55	3.58
Augmented coref.	26.73	5.40
Global information	31.76	0.37

Table 3: Results of ablation tests relative to *BIU-DISC_{no-loc}*. The columns specify the component removed, the micro-averaged F_1 score achieved without it, and the marginal contribution of the component.

nor difference, which is not surprising considering the conservative approach we took, using a single global term for each sentence. Possibly, this information is also included in the other components, thus proving no marginal contribution relative to them. Under the conditions of an overwhelming majority of negative examples, this is a risky method to use, and should be considered when the ratio of positive examples is higher. For future work, we intend to use this information via classification features (e.g. the coverage obtained with and without global information), rather than the crude addition of the term to the sentence.

Analysis of augmented coreferences We analyzed the performance of the component for augmenting coreference relations relative to the OpenNLP resolver. Recall that our component works on top of the resolver’s output and can add or remove coreference relations. As a complete annotation of coreference chains in the dataset is unavailable, we performed the following evaluation. Recall is computed based on the number of identified pairs from a sample of 100 intra-document coreference and bridging relations from the annotated dataset described in (Mirkin et al., 2010). Precision is computed based on 50 pairs sampled from the output of each method, equally distributed over topics. The results, shown in Table 4, indicate the much higher recall obtained by our component at some cost in precision. Although rather simple, the ablation test of this component shows its usefulness. Still, both methods achieve low absolute recall, suggesting the need for more robust tools for this task.

	P (%)	R (%)	F_1 (%)
OpenNLP	74	16	26.3
Augmented coref.	60	28	38.2

Table 4: Performance of coreference methods.

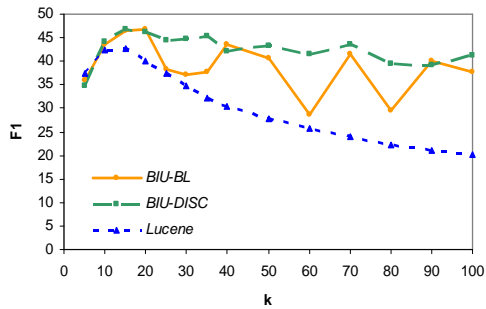


Figure 3: F_1 performance as a function of the number of retrieved candidates.

Candidate retrieval setting As mentioned in Section 3, best performance of RTE systems in the task was obtained when applying a first step of IR-based candidate filtering. We therefore compare the performance of *BIU-DISC* with that of *BIU-BL* under this setting as well.⁴ For candidate retrieval we used Lucene, a state of the art search engine⁵, in a range of top- k retrieved candidates. The results are shown in Figure 3. For reference, the figure also shows the performance along this range of Lucene as-is, when no further inference is applied to the retrieved candidates.

While *BIU-DISC* does not outperform *BIU-BL* at every point, the area under the curve is clearly larger for *BIU-DISC*. The figure also indicates that *BIU-DISC* is far more robust, maintaining a stable F_1 and enabling a stable tradeoff between recall and precision along the whole range (recall ranges between 42% and 55% for $k \in [15 - 100]$, with corresponding precision range of 51% to 33%).

Finally, Table 5 shows the results of the best systems as determined in our first experiment. We performed a single experiment to compare *BIU-DISC_{no-loc}* and *BIU-BL*³ under a candidate retrieval setting, using $k = 20$, where both systems highly perform. We compare these results to the highest score obtained by Lucene, as well as to the two best submissions to the RTE-5 Search task⁶. *BIU-DISC_{no-loc}* outperforms all other methods and its result is significantly better than *BIU-BL*³ with $p < 0.01$ according to McNemar’s test.

⁴This time, for global information, the document’s three highest ranking terms were added to each sentence.

⁵<http://lucene.apache.org>

⁶The best one is an earlier version of this work (Mirkin et al., 2009); the second is MacKinlay and Baldwin’s (2009).

	P (%)	R (%)	F_1 (%)
<i>BIU-DISC_{no-loc}</i>	50.77	45.12	47.78
<i>BIU-BL</i> ³	51.68	40.38	45.33
Lucene, top-15	35.93	52.50	42.66
RTE-5 best	40.98	51.38	45.59
RTE-5 second-best	42.94	38.00	40.32

Table 5: Performance of best configurations.

7 Conclusions

While it is generally assumed that discourse interacts with semantic entailment inference, the concrete impacts of discourse on such inference have been hardly explored. This paper presented a first empirical investigation of discourse processing aspects related to entailment. We argue that available discourse processing tools should be substantially improved towards this end, both in terms of the phenomena they address today, namely nominal coreference, and with respect to the covering of additional phenomena, such as bridging anaphora. Our experiments show that even rather simple methods for addressing discourse can have a substantial positive impact on the performance of entailment inference. Concerning the locality phenomenon stemming from discourse coherence, we learned that it does carry potentially useful information, which might become beneficial in the future when better-performing entailment systems become available. Until then, integrating this information with entailment confidence may be useful. Overall, we suggest that entailment systems should extensively incorporate discourse information, while developing sound algorithms for addressing various discourse phenomena, including the ones described in this paper.

Acknowledgements

The authors are thankful to Asher Stern and Ilya Kogan for their help in implementing and evaluating the augmented coreference component, and to Roy Bar-Haim for useful advice concerning this paper. This work was partially supported by the Israel Science Foundation grant 1112/08 and the PASCAL-2 Network of Excellence of the European Community FP7-ICT-2007-1-216886. Jonathan Berant is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship.

References

- Bar-Haim, Roy, Jonathan Berant, Ido Dagan, Ido Greental, Shachar Mirkin, Eyal Shnarch, and Idan Szpektor. 2008. Efficient semantic deduction and approximate matching over compact parse forests. In *Proc. of Text Analysis Conference (TAC)*.
- Bar-Haim, Roy, Jonathan Berant, and Ido Dagan. 2009. A compact forest for scalable inference over entailment and paraphrase rules. In *Proc. of EMNLP*.
- Bensley, Jeremy and Andrew Hickl. 2008. Unsupervised resource creation for textual inference applications. In *Proc. of LREC*.
- Bentivogli, Luisa, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, Medea Lo Leggio, and Bernardo Magnini. 2009a. Considering discourse references in textual entailment annotation. In *Proc. of the 5th International Conference on Generative Approaches to the Lexicon (GL2009)*.
- Bentivogli, Luisa, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009b. The fifth PASCAL recognizing textual entailment challenge. In *Proc. of TAC*.
- Castillo, Julio J. 2009. Sagan in TAC2009: Using support vector machines in recognizing textual entailment and TE search pilot task. In *Proc. of TAC*.
- Clark, Peter and Phil Harrison. 2009. An inference-based approach to recognizing entailment. In *Proc. of TAC*.
- Clark, Herbert H. 1975. Bridging. In Schank, R. C. and B. L. Nash-Webber, editors, *Theoretical issues in natural language processing*, pages 169–174. Association of Computing Machinery.
- Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognizing textual entailment challenge. In *Machine Learning Challenges*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.
- Dagan, Ido, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, pages 15(4):1–17.
- Dali, Lorand, Delia Rusu, Blaz Fortuna, Dunja Mladenic, and Marko Grobelnik. 2009. Question answering based on semantic graphs. In *Proc. of the Workshop on Semantic Search (SemSearch 2009)*.
- Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.
- Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. of ACL*.
- Giampiccolo, Danilo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proc. of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.
- Harabagiu, Sanda and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proc. of ACL*.
- Harman, Donna. 1992. The DARPA TIPSTER project. *SIGIR Forum*, 26(2):26–28.
- Joachims, Thorsten. 2006. Training linear SVMs in linear time. In *Proc. of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.
- Li, Fangtao, Yang Tang, Minlie Huang, and Xiaoyan Zhu. 2009. Answering opinion questions with random walks on graphs. In *Proc. of ACL-IJCNLP*.
- Litkowski, Ken. 2009. Overlap analysis in textual entailment recognition. In *Proc. of TAC*.
- MacKinlay, Andrew and Timothy Baldwin. 2009. A baseline approach to the RTE5 search pilot. In *Proc. of TAC*.
- Mirkin, Shachar, Roy Bar-Haim, Jonathan Berant, Ido Dagan Eyal Shnarch, Asher Stern, and Idan Szpektor. 2009. Addressing discourse and document structure in the RTE search task. In *Proc. of TAC*.
- Mirkin, Shachar, Ido Dagan, and Sebastian Padó. 2010. Assessing the role of discourse references in entailment inference. In *Proc. of ACL*.
- Nenkova, Ani, Rebecca Passonneau, and Kathleen Mckeown. 2007. The pyramid method: incorporating human content selection variation in summarization evaluation. In *ACM Transactions on Speech and Language Processing*.
- Qiu, Long, Min-Yen Kan, and Tat-Seng Chua. 2004. A public reference implementation of the RAP anaphora resolution algorithm. In *Proc. of LREC*.
- Romano, Lorenza, Milen Kouylekov, Idan Szpektor, and Ido Dagan. 2006. Investigating a generic paraphrase-based approach for relation extraction. In *Proc. of EACL*.
- Witten, Ian H. and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco.