Assessing the Role of Discourse References in Entailment Inference

Shachar Mirkin, Ido Dagan
Bar-Ilan University
Ramat-Gan, Israel
{mirkins,dagan}@cs.biu.ac.il

Sebastian Padó
University of Stuttgart
Stuttgart, Germany
pado@ims.uni-stuttgart.de

Abstract

Discourse references, notably coreference and bridging, play an important role in many text understanding applications, but their impact on textual entailment is yet to be systematically understood. On the basis of an in-depth analysis of entailment instances, we argue that discourse references have the potential of substantially improving textual entailment recognition, and identify a number of research directions towards this goal.

1 Introduction

The detection and resolution of *discourse references* such as coreference and bridging anaphora play an important role in text understanding (TU) applications, like Question Answering (QA) and Information Extraction (IE). There, reference resolution is used for the purpose of combining knowledge from multiple sentences.

Such knowledge is arguably also important for Textual Entailment (TE), a generic framework for modeling semantic inference. TE reduces the inference requirements of many TU applications to the problem of determining whether the meaning of a certain *text* (T) can be inferred from the meaning of a given textual assertion, termed *hypothesis* (H) (Dagan et al., 2006).

Consider the following example:

(1) T: "*Not only had he developed an aversion to the* ***President***$_1$ *and politics in general,* ***Oswald***$_2$ *was also a failure with Marina, his wife. [...] Their relationship was supposedly responsible for why* ***he***$_2$ *killed* ***Kennedy***$_1$."

H: "*Oswald killed President Kennedy.*"

The understanding that the second sentence of the text entails the hypothesis draws on two coreference relationships, namely that *he* is *Oswald*, and that the *Kennedy* in question is *President Kennedy*. However, the utilization of discourse information has been so far limited mainly to the substitution of nominal coreferents, while many aspects of the interface between discourse and the semantic needs of such applications remain largely unknown.

The recently held Fifth Recognizing Textual Entailment (RTE-5) challenge (Bentivogli et al., 2009a) has introduced a *Search* task, where the text sentences are interpreted in the context of their full discourse, as in Example 1 above. Accordingly, TE constitutes an interesting framework – and the Search task an adequate dataset – to study the interrelation between discourse and inference.

The goal of this study is to analyze the roles of discourse references for textual entailment inference, to provide relevant findings and insights to developers of both reference resolvers and entailment systems and to highlight promising directions for the better incorporation of discourse phenomena into inference. Our focus is on a manual, in-depth assessment that results in a classification and quantification of discourse reference phenomena and their utilization for inference. On this basis, we develop an account of formal devices for incorporating discourse references into the inference computation. An additional point of interest is the interrelation between entailment knowledge and coreference. E.g., in Example 1 above, knowing that Kennedy was a president can alleviate the need for coreference resolution. Conversely, coreference resolution can often be used to overcome gaps in entailment knowledge.

Structure of the paper. In Section 2, we provide background on the use of discourse references in natural language processing (NLP) in general and specifically in TE. Section 3 describes

the goals of this study, followed by our analysis scheme (Section 4) and the required inference mechanisms (Section 5). Section 6 presents quantitative results and further observations. Conclusions and further research are discussed in Section 7.

2 Background

2.1 Discourse in NLP

Discourse information plays a role in a range of NLP tasks. It is obviously central to *discourse processing* tasks such as text segmentation (Hearst, 1997). Reference information provided by discourse is also useful for *text understanding* tasks such as QA, IE, and IR (Vicedo and Ferrndez, 2006; Zelenko et al., 2004; Na and Ng, 2009), as well as for the acquisition of lexical-semantic "narrative schema" knowledge (Chambers and Jurafsky, 2009). Discourse references have been the subject of attention in both the Message Understanding Conference (Grishman and Sundheim, 1996) and the Automatic Content Extraction program (Strassel et al., 2008).

The simplest form of information that discourse provides is *coreference*, i.e., information that two linguistic expressions refer to the same entity or event. Coreference is particularly important for processing pronouns and other anaphoric expressions, such as *he* in Example 1. Ability to resolve this reference translates directly into, e.g., a QA system's ability to answer questions like *Who killed Kennedy?*.

A second, more complex type of information stems from *bridging references*, such as in the following discourse (Asher and Lascarides, 1998):

(2) "*I've just **arrived**. The **camel** is outside.*"

While coreference indicates equivalence, bridging points to the existence of a salient semantic relation between two distinct entities or events. Here, it is (informally) *'means of transport'*, which would make the discourse (2) relevant for a question like *How did I arrive here?*. Other types of bridging relations include set-membership, roles in events, causation and consequence (Clark, 1975).

Note, however, that practical text understanding systems are generally limited to the resolution of entity (or even just pronoun) coreference (Li et al., 2009; Dali et al., 2009). One important reason is the unavailability of tools to resolve the more complex (and difficult) forms of discourse reference such as event coreference and bridging.[1] Another one is uncertainty about their practical importance.

2.2 Discourse in Textual Entailment

Textual Entailment has been introduced in Section 1 as a common-sense notion of inference. It has spawned interest in the computational linguistics community as a common denominator of many NLP tasks including IE, summarization and tutoring (Romano et al., 2006; Harabagiu et al., 2007; Nielsen et al., 2009).

Architectures for Textual Entailment. Over the course of recent RTE challenges (Giampiccolo et al., 2007; Giampiccolo et al., 2008), the main benchmark for TE technology, two architectures for modeling TE have emerged as dominant: *transformations* and *alignment*. The goal of ***transformation***-based TE models is to determine the entailment relation $T \Rightarrow H$ by finding a "proof", i.e., a sequence of consequents, $(T, T_1, \ldots, T_n)$, such that T_n=H (Bar-Haim et al., 2008; Harmeling, 2009), and that in each transformation, $T_i \rightarrow T_{i+1}$, the consequent T_{i+1} is entailed by T_i. These transformations commonly include lexical modifications and the generation of syntactic alternatives. The second major approach constructs an *alignment* between the linguistic entities of the trees (or graphs) of T and H, which can represent syntactic structure, semantic structure, or non-hierarchical phrases (Zanzotto et al., 2009; Burchardt et al., 2009; MacCartney et al., 2008). H is assumed to be entailed by T if its entities can be aligned "well" to corresponding entities in T. Alignment quality is generally determined based on features that assess the validity of the local replacement of the T entity by the H entity.

While transformation- and alignment-based entailment models look different at first glance, they ultimately have the same goal, namely obtaining a maximal *coverage* of H by T, i.e. to identify matches of as many elements of H within T as possible. [2] To do so, both architectures typically make use of *inference rules* such as '*Y was pur-*

[1] Some works, e.g. (Markert et al., 2003; Poesio et al., 2004), address the resolution of few specific kinds of bridging references, such as meronymic ones; yet, wide-scope systems for bridging resolutions are unavailable.

[2] Clearly, the details of how the final entailment decision is made based on the attained coverage differ substantially among models.

chased by X → X paid for Y', either by directly applying them as transformations, or by using them to score alignments. Rules are generally drawn from external knowledge resources, such as WordNet (Fellbaum, 1998) or DIRT (Lin and Pantel, 2001), although knowledge gaps remain a key obstacle (Bos, 2005; Balahur et al., 2008; Bar-Haim et al., 2008).

Discourse in previous RTE challenges. The first instances of the RTE challenge used "self-contained" texts and hypotheses, where discourse considerations played virtually no role. A first step towards a more comprehensive notion of entailment was taken with RTE-3 (Giampiccolo et al., 2007), when paragraph-length texts were first included and constituted 17% of the texts in the test set. Chambers et al. (2007) report that in a sample of $T - H$ pairs drawn from the development set, 25% involved discourse references.

Using the concepts introduced above, the impact of discourse references can be generally described as a *coverage problem*, independent of the system's architecture. In Example 1, the hypothesis word *Oswald* cannot be safely linked to the text pronoun *he* without further knowledge about *he*; the same is true for '*Kennedy → President Kennedy*' which involves a specialization that is only warranted in the specific discourse.

A number of systems have tried to address the question of coreference in RTE as a preprocessing step prior to inference proper, with most systems using off-the-shelf coreference resolvers such as JavaRap (Qiu et al., 2004) or OpenNLP[3]. Generally, anaphoric expressions were textually replaced by their antecedents. Results were inconclusive, however, with several reports about errors introduced by automatic coreference resolution (Agichtein et al., 2008; Adams et al., 2007). Specific evaluations of the contribution of coreference resolution yielded both small negative (Bar-Haim et al., 2008) and insignificant positive (Chambers et al., 2007) results.

3 Motivation and Goals

The results of recent studies, as reported in Section 2.2, seem to show that the resolution of discourse references in RTE systems hardly affects performance. However, our intuition is that these results can be attributed to four major limitations shared by these studies: (1) the datasets, where discourse phenomena were not well represented; (2) the off-the-shelf coreference resolution systems which may have been not robust enough; (3) the limitation to nominal coreference; and (4) overly simple integration of reference information into the inference engines.

The goal of this paper is to assess the impact of discourse references on entailment with an annotation study which removes these limitations. To counteract (1), we use the recent RTE-5 Search dataset (details below). To avoid (2), we perform a manual analysis, assuming discourse references as predicted by an oracle. With regards to (3), our annotation scheme covers coreference and bridging relations of all syntactic categories and classifies them. As for (4), we suggest several operations necessary to integrate the discourse information into an entailment engine.

In contrast to the numerous existing datasets annotated for discourse references (Hovy et al., 2006; Strassel et al., 2008), we do not annotate exhaustively. Rather, we are interested specifically in those instances of references that impact inference. Furthermore, we analyze each instance from an entailment perspective, characterizing the relevant factors that have an impact on inference. To our knowledge, this is the first such in-depth study.[4].

The results of our study are of twofold interest. First, they provide guidance for the developers of reference resolvers who might prioritize the scope of their systems to make them more valuable for inference. Second, they point out potential directions for the developers of inference systems by specifying what additional inference mechanisms are needed to utilize discourse information.

The RTE-5 Search dataset. We base our annotation on the Search task dataset, a new addition to the recent Fifth RTE challenge (Bentivogli et al., 2009a) that is motivated by the needs of NLP applications and drawn from the TAC summarization track. In the Search task, TE systems are required to find *all* individual sentences in a given corpus which entail the hypothesis – a setting that is sensible not only for summarization, but also for information access tasks like QA. Sentences are judged individually, but "are to be interpreted in the context of the corpus as they rely on explicit

[3] http://opennlp.sourceforge.net

[4] The guidelines and the dataset are available at http://www.cs.biu.ac.il/~nlp/downloads

and implicit references to entities, events, dates, places, etc., mentioned elsewhere in the corpus" (Bentivogli et al., 2009b).

4 Analysis Scheme

For annotating the RTE-5 data, we operationalize reference relations that are *relevant* for entailment as those that improve *coverage*. Recall from Section 2.2 that the concept of coverage is applicable to both transformation and alignment models, all of which aim at maximizing coverage of H by T.

We represent T and H as syntactic trees, as common in the RTE literature (Zanzotto et al., 2009; Agichtein et al., 2008). Specifically, we assume MINIPAR-style (Lin, 1993) dependency trees where nodes represent text expressions and edges represent the syntactic relations between them. We use "term" to refer to text expressions, and "components" to refer to nodes, edges, and subtrees. Dependency trees are a popular choice in RTE since they offer a fairly semantics-oriented account of the sentence structure that can still be constructed robustly. In an ideal case of entailment, all nodes *and* dependency edges of H are covered by T.

For each $T - H$ pair, we annotate all relevant discourse references in terms of three textual items: the *target component* in H, the *focus term* in T, and the *reference term* which stands in a reference relation to the focus term. By resolving this reference, the target component can usually be inferred; sometimes, however, more than one reference term needs to be found. We now define and illustrate these concepts on examples from Table 1. [5]

The ***target component*** is a tree component in H that cannot be covered by the "local" material from T. An example for a tree component is Example (v), where the target component *AS-28 mini submarine* in H cannot be inferred from the pronoun *it* in T. Example (vi) demonstrates an edge as target component. In this case, the edge in H connecting *melt* with the modifier *in the Arctic* is not found in T. Although each of the hypothesis' nodes can be covered separately via knowledge-based rules (e.g. '*Siberia* $\rightarrow$ *Arctic*', '*permafrost* $\rightarrow$ *ice*', '*thaw* $\leftrightarrow$ *melt*'), the resulting fragments in T are unconnected without the (intra-sentential) coreference *them* and *lakes in Siberia*.

For each target component, we identify its ***focus term*** as the expression in T that does not cover the target component itself but participates in a reference relation that can help covering it.

We follow the focus term's reference chain to a ***reference term*** which can, either separately or in combination with the focus term, help covering the target component. In Example (ii), where the target component in H is *2003 UB313*, *Xena* is the focus term in T and the reference term is a mention of *2003 UB313* in a previous sentence, T'. In this case, the reference term covers the entire target component on its own.

An additional attribute that we record for each instance is whether resolving the discourse reference is *mandatory* for determining entailment, or *optional*. In Example (v), it is mandatory: the inference cannot be completed without the knowledge provided by the discourse. In contrast, in Example (ii), inferring *2003 UB313* from *Xena* is optional. It can be done either by identifying their coreference relation, or by using background knowledge in the form of an entailment rule, '*Xena* $\leftrightarrow$ *2003 UB313*', that is applicable in the context of astronomy. Optional discourse references represent instances where discourse information and TE knowledge are interchangeable. As mentioned, knowledge gaps constitute a major obstacle for TE systems, and we cannot rely on the availability of any ceratin piece of knowledge to the inference process. Thus, in our scheme, mandatory references provide a "lower bound" with regards to the necessity to resolve discourse references, even in the presence of *complete knowledge*; optional references, on the other hand, set an "upper bound" for the contribution of discourse resolution to inference, when *no knowledge* is available. At the same time, this scheme allows investigating how much TE knowledge can be replaced by (perfect) discourse processing.

When choosing a reference term, we search the reference chain of the focus term for the nearest expression that is identical to the target component or a subcomponent of it. If we find such an expression, covering the identical part of the target component requires no entailment knowledge. If no identical reference term exists, we choose the semantically 'closest' term from the reference chain, i.e. the term which requires the least knowledge

[5] In our annotation, we assume throughout that some basic knowledge about admissible syntactic transformations is available, such as passive to active or derivational transformations; for brevity, we ignore articles in the examples and treat named entities as single nodes.

	Text		Hypothesis
i	T'	*Once the reform becomes law, Spain will join the Netherlands and Belgium in allowing* ***homosexual marriages.***	*Massachusetts allows* ***gay marriages***
	T	***Such unions*** *are also legal in six Canadian provinces and the northeastern US state of Massachusetts.*	
ii	T'	*The official name of* ***2003 UB313*** *has yet to be determined.*	***2003 UB313*** *is in the Kuiper Belt*
	T	*Brown said he expected to find a moon orbiting* ***Xena*** *because many Kuiper Belt objects are paired with moons.*	
iii	T'_a	*All seven aboard the* ***AS-28 submarine*** *appeared to be in satisfactory condition, naval spokesman said.*	*The* ***AS-28 mini submarine*** *was trapped underwater*
	T'_b	*British crews were working with Russian naval authorities to maneuver the unmanned robotic vehicle and untangle the* ***AS-28****.*	
	T	*The Russian military was racing against time early Friday to rescue a* ***mini submarine trapped on the seabed.***	
iv	T'	***China*** *seeks solutions to its coal mine safety.*	*A* ***mining accident in China*** *has killed several miners.*
	T	*A* ***recent accident*** *has cost more than a dozen miners their lives.*	
v	T''	*A remote-controlled device was lowered to the stricken vessel to cut the cables in which the* ***AS-28 vehicle*** *is caught.*	*The* ***AS-28 mini submarine*** *was trapped underwater*
	T'	*The* ***mini submarine*** *was resting on the seabed at a depth of about 200 meters.*	
	T	*Specialists said* ***it*** *could have become tangled up with a metal cable or in sunken nets from a fishing trawler.*	
vi	T	*. . . dried up* ***lakes in Siberia****, because the permafrost beneath* ***them*** *has begun to thaw.*	*The ice is* ***melting in the Arctic***

Table 1: Examples for discourse-dependent entailment in the RTE-5 dataset, where the inference of H depends on reference information from the discourse sentences T' / T''. Referring terms (in T) and target terms (in H) are shown in boldface.

to infer the target component. For instance, suppose the target is *ice* and the focus term is *it*, if we cannot find *ice* coreferring with the focus term, we pick the semantically closest term coreferring with it, e.g. *permafrost*.

Finally, for each reference relation that we annotate, we record four additional attributes which we assumed to be informative in an evaluation. First, the *reference type*: Is the relation a coreference or a bridging reference? Second, the *syntactic type* of the focus and reference terms. Third, the *focus/reference terms entailment status* – does some kind of entailment relation hold between the two terms? Fourth, the *operation* that should be performed on the focus and reference terms to obtain coverage of the target component (see Section 5).

5 Integrating Discourse References into Entailment Recognition

The result of our initial analysis was that the standard substitution operation applied by virtually all previous studies for integrating coreference into entailment is insufficient. We identified a total of three distinct cases for the integration of discourse reference knowledge for entailment, which correspond to different relations between the target component, the focus term and the reference term. This section describes the three cases and characterizes them in terms of tree transformations. We assume a transformation-based entailment architecture (cf. Section 2.2), although we believe that the key points of our account are also applicable to alignment-based architecture. Transformations create revised trees that cover previously uncovered target components in H. The output of each transformation, T_1, is comprised of copies of the components used to construct it, and is appended to the discourse forest, which includes the dependency trees of all sentences and their generated consequents. When a component is cloned, the original and cloned nodes are connected with coreference links. [6]

We assume that we have access to a dependency tree for H, a dependency forest for T and its discourse context, as well as the output of a perfect discourse processor, i.e., a complete set of both coreference and bridging relations, including the type of bridging relation (e.g. *part-of*, *member-of*, *cause*).

We use the following notation. We use x, y for tree nodes, and S_x to denote a (sub-)tree with root x. $lab(x)$ is the label of the incoming edge of x (i.e., its grammatical function). We write

[6] Hypotheses in the RTE datasets rarely contain discourse references. Nevertheless, an analysis of the RTE-5 hypotheses shows that this procedure is also applicable to H, if necessary.

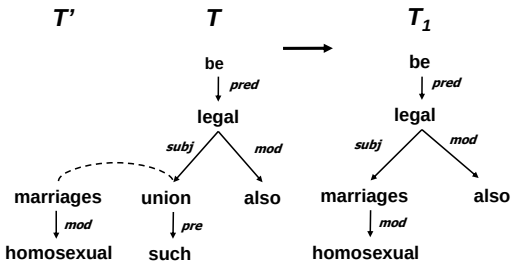

Figure 1: The *Substitution* transformation, demonstrated on the relevant subtrees of Example (i). The dashed line denotes a discourse reference.

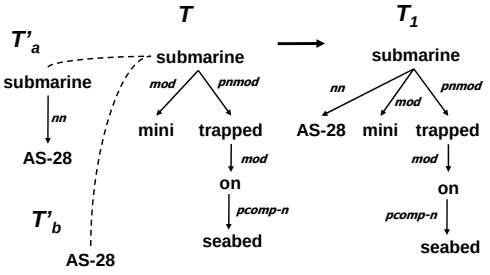

Figure 2: The *dependent-merge* (T'_a) and *head-merge* (T'_b) transformations (Example (iii)).

$C(x, y)$ for a coreference relation between S_x and S_y, the corresponding trees of the focus and reference terms, respectively. We write $B_r(x, y)$ for a bridging relation, where r is its type.

(1) Substitution: This is the most intuitive and widely used transformation, which corresponds directly to the treatment of discourse information in existing systems. It applies to coreference relations, when an expression found elsewhere in the text (the reference term) can cover all missing information (the target component) on its own. In such cases, the reference term can replace the entire focus term. Apparently (cf. Section 6), substitution applies also to some types of bridging relations, such as *set-membership*, when the member is sufficient for representing the entire set for the necessary inference. For example, in *I met two people yesterday. The woman told me a story* (Clark, 1975), substituting *two people* with *woman* results in a text which is entailed from the discourse and may be useful for inference.

In a parse tree representation, given a coreference relation $C(x, y)$ (or $B_r(x, y)$, respectively), the newly generated tree, T_1, consists of a copy of T, where the entire tree S_x is replaced by a copy of S_y . In Figure 1, which shows Example (i) from Table 1, *such unions* is substituted by *homosexual marriages*.

Head-substitution Occasionally, substituting only the head of the focus term is sufficient. In such cases, only the root nodes x and y are substituted. This is the case, for example, with synonymous verbs (e.g. *melt* & *thaw*). As verbs typically constitute tree roots in dependency parses, substituting or merging (see below) their entire trees might be wasteful. In such cases, the simpler head-substitution may be applied.

(2) Merge: In contrast to substitution, where a match for the entire target component is found elsewhere in the text, this transformation is required when parts of the missing information are scattered among multiple locations in the text. We distinguish between two types of merge transformations: (a) *dependent-merge*, and (b) *head-merge*, depending on the syntactic roles of the merged components.

(a) Dependent-Merge In a dependent-merge operation the head of (at least) one of the merged terms covers the head node of the target component, but modifiers from both of them are required to cover the target component's dependent nodes. In our representation, we need to place copies of all dependents of coreferring trees under one single root node in T_1.

The transformation is illustrated in Figure 2, using Example (iii). The component we wish to cover in H is the noun phrase *AS-28 mini submarine*. Unfortunately, the focus term in T, "*mini submarine trapped on the seabed*", covers only the modifier *mini*, but not *AS-28*. This modifier can however be provided by the coreferent term in T'_a (left upper corner). Now, the inference engine can, e.g., employ the rule '*on seabed* $\rightarrow$ *underwater*' to cover H completely.

Suppose, without loss of generality, that y matches the root node of the target component. Formally, given $C(x, y)$, we define T_1 as a copy of T, where (i) the subtree S_x is replaced by S_y, and (ii) for all children c of x, a copy of S_c is placed under the copy of y in T_1 with its original edge label, $lab(c)$. If x is the node that matches the root of the target component, the children y will be placed as dependents of S_x in T_1.

(b) Head-merge An alternative way to recover the missing information in Example (iii) is to find a reference term whose head word itself (rather

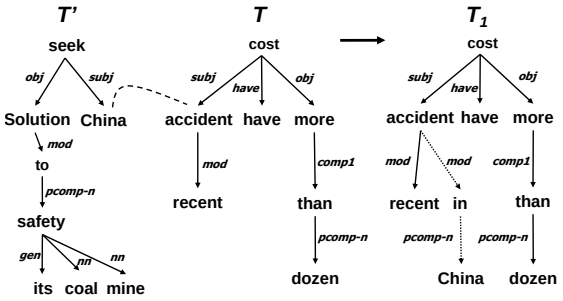

Figure 3: The *insertion* transformation. Dotted edges mark the newly inserted path (Example (iv)).

than one of its modifiers) is the target component's missing modifier, as shown in Figure 2 in the bottom left corner (T'_b).

In terms of parse trees, we need to add one tree, as a dependent of the other. Formally, given $C(x, y)$, similarly to dependent-merge, T_1 is created as a copy of T where the subtree S_x is replaced by either S_x or S_y, depending on whichever of x and y matches the target component's head. Assume it is x, for example. Then, a copy of S_y is added as a new child to x with the edge label nn (nominal modifier). This choice of label is motivated by the fact that cases of this kind typically correspond to internal coreferences within nominal target components (such as between *AS-28* and *mini submarine* in this case). Another example is the combination of *the coastal city* (focus term) and *Karachi* (reference term) into *The coastal city of Karachi*).

(3) Insertion: The last transformation, Insertion, is used when a relation that is realized in H is missing from T and is only implied via a bridging relation. Consider Example (iv). as shown in Figure 3, the location that is explicitly mentioned in H can only be covered by T by resolving a bridging reference with *China* in T'. Formally, given a bridging relation $B_r(x, y)$, we introduce a new subtree S^r_z into T_1, where z is a child of x and $lab(z) = lab_r$. S^r_z must have a unique *gap*, g, that is filled with a copy of $S(y)$.

This transformation stands out from the others in that it introduces new material. For each bridging relation, it adds a specific subtrees S^r via an edge labeled with lab_r. These two items form the dependency representation of the bridging relation B_r and must be provided by the interface between the discourse and the inference systems. Clearly, their exact form depends on the set of bridging relations provided by the discourse resolver as well as the details of the dependency parses.

In our example, the bridging relation *located-in* (r) is represented by inserting a subtree S^r_z headed by *in* (z) into T_1 and connecting it to *accident* (x) as a modifier (lab_r). The subtree S^r_z consists of a gap g which is connected to *in* with a *pcomp-n* dependency (a nominal head of a prepositional phrase), and which is filled with the node *China* (y) when the transformation is applied. Note that the structure of S^r_z and the way it is inserted into T_1 are predefined by the above-mentioned interface; only the node to which it is attached and the contents of g are determined at transformation-time.

As another example, consider the following short text from (Clark, 1975): *John was murdered yesterday. The knife lay nearby*. Here, the bridging relation between the murder event and the instrument, *the knife* (x), can be addressed by inserting under x a subtree for the clause *with which* as S^r_z, with a gap which is filled by the prase-tree (headed by *murdered*, y) of entire first sentence *John was murdered yesterday*.

Transformation chaining. Since our transformations are defined to be minimal, some cases require the application of multiple transformations to achieve coverage. Consider Example (v), Table 1. We wish to cover *AS-28 mini submarine* in H from the coreferring *it* in T, *mini submarine* in T' and *AS-28 vehicle* in T''. A substitution of *it* by either coreference does not suffice, since none of the antecedents contains all necessary modifiers. It is therefore necessary to substitute *it* first by one of the coreferences and then merge it with the other.

6 Results

We analyzed 120 sentence-hypothesis pairs of the RTE-5 development set (21 different hypotheses, 111 distinct sentences, 53 different documents). Below, we summarize our findings, focusing on the relation between our findings and the assumptions of previous studies as discussed in Section 3.

General statistics. We found that 44% of the pairs contained reference relations whose resolution was mandatory for inference. In another 28%, references could optionally support the inference of the hypothesis. In the remaining 28%, references did not contribute towards inference. The

(%)	Pronoun	NE	NP	VP
Focus term	9	19	49	23
Reference term	-	43	43	14

Table 2: Syntactic types of discourse references

(%)	Sub.	Merge	Insertion
Coreference	62	38	-
Bridging	30	-	70
Total	54	28	18

Table 3: Distribution of transformation types

total number of relevant references was 137, and 37 pairs (27%) contained multiple relevant references. These numbers support our assumption that discourse references play an important role in inference.

Reference types. 73% of the identified references are coreferences and 27% are bridging relations. The most common bridging relation was the location of events (e.g. *Arctic* in Arctic ice melting events), generally assumed to be known throughout the document. Other bridging relations we encountered included cause (e.g. between *injured* and *the attack*), event participants and set membership.

Syntactic types. Table 2 shows that roughly three quarters of all focus and reference terms were nominal phrases, which justifies their prominent position in work on anaphora and coreference resolution. However, 23% of the focus terms were verbal phrases. We found these focus terms to be frequently crucial for entailment since they included the main predicate of the hypothesis.[7] This calls for an increased focus on the resolution of event references.

Transformations. Table 3 shows the relative frequencies of all transformations. Again, we found that the "default" transformation, substitution, is the most frequent one, and is helpful for both coreference and bridging relations. Substitution is particularly useful for handling pronouns (14% of all substitution instances), the replacement of named entities by synonymous names (32%), the replacement of other NPs (38%), and the substitution of verbal head nodes in event coreference (16%). Yet, in nearly half the cases, a different transformation had to be applied. Insertion accounts for the majority of bridging cases. Head-merge is necessary to integrate proper nouns as modifiers of other head nouns from H. Dependent-merge, responsible for 85% of the merge transformations, can be used to complete nominal focus terms with missing modifiers (e.g., adjectives), as well as for merging other dependencies between coreferring predicates. This result indicates the importance of incorporating other transformations into inference systems.

Distance of reference terms. The distance between the focus and the reference terms varied considerably, ranging from intra-sentential reference relations and up to several dozen sentences. For more than a quarter of the focus terms, we had to go to other documents to find matches for the target components that, in conjunction with the focus term, could cover the target component. Interestingly, all such cases involved coreference (about equally divided between the merge transformations and substitutions), while bridging was always "document-local". This result reaffirms the usefulness of cross-document coreference resolution for inference (Huang et al., 2009).

Discourse resolution as preprocessing? In existing RTE systems, discourse references are typically resolved as a preprocessing step. While our annotation was manual and cannot yield direct results about processing considerations, we observed that discourse relations often hold between complex, and deeply embedded, expressions, which makes their automatic resolution difficult. Of course, many RTE systems attempt to normalize and simplify H and T, e.g., by splitting conjunctions or removing irrelevant clauses, but these operations are usually considered a part of the inference rather the preprocessing phase (cf. e.g., Bar-Haim et al. (2007)). Since the resolution of discourse references is likely to profit from these steps, it seems desirable to "postpone" it until after simplification. In transformation-based systems, it might be natural to add discourse-based transformations to the set of inference operations, while in alignment-based systems, discourse references can be integrated into the computation of alignment scores.

Discourse references vs. entailment knowledge. We have stated before that even if a discourse reference is not strictly necessary for entailment, it may be interesting because it represents an alter-

[7] The lower proportion of VPs among reference terms stems from bridging relations between VPs and nominal dependents, such as the abovementioned "location" relation.

native to the use of knowledge rules to cover the hypothesis. Sometimes, these rules are generally applicable (e.g., '*Alaska* $\rightarrow$ *Arctic*'). However, often they are context-specific. Consider the following sentence as T for the H "*The ice is melting in the Arctic*":

(3) T: "*The scene at the* ***receding*** *edge of the Exit Glacier was part festive gathering, part nature tour with an apocalyptic edge.*"

While it is possible to cover *melting* using a rule '*melting* $\leftrightarrow$ *receding*', this rule is only valid under quite specific conditions (e.g., for the subject *ice*). Instead of determining the applicability of the rule, a discourse-aware system can take the next sentence into account, which contains a coreferring event to *receding* that can cover *melting* in H:

(4) T': "*... people moved closer to the rope line near the glacier as it shied away, practically groaning and* ***melting*** *before their eyes.*"

Discourse relations can in fact encode arbitrarily complex world knowledge, as in the following pair:

(5) H: "*The* ***serial*** *killer BTK was accused of at least 7 killings starting in the 1970's.*"

T: "*Police say BTK may have killed as many as 10 people between 1974 and 1991.*"

Here, the H modifier *serial*, which does not occur in T, can be covered either by world knowledge (a person who killed 10 people is a serial killer), or by resolving the coreference of *BTK* to the term *the serial killer BTK* which occurs in the discourse around T. Our conclusion is that not only can discourse references often replace world knowledge in principle, in practice it often seems easier to resolve discourse references than to determine whether a rule is applicable in a given context or to formalize complex world knowledge as inference rules. Our annotation provides further empirical support to this claim: An entailment relation exists between the focus and reference terms in 60% of the focus-reference term pairs, and in many of the remainder, entailment holds between the terms' heads. Thus, discourse provides relations which are many times equivalent to entailment knowledge rules and can therefore be utilized in their stead.

7 Conclusions

This work has presented an analysis of the relation between discourse references and textual entailment. We have identified a set of limitations common to the handling of discourse relations in virtually all entailment systems. They include the use of off-the-shelf resolvers that concentrate on nominal coreference, the integration of reference information through substitution, and the RTE evaluation schemes, which played down the role of discourse. Since in practical settings, discourse plays an important role, our goal was to develop an agenda for improving the handling of discourse references in entailment-based inference.

Our manual analysis of the RTE-5 dataset shows that while the majority of discourse references that affect inference are nominal coreference relations, another substantial part is made up by verbal terms and bridging relations. Furthermore, we have demonstrated that substitution alone is insufficient to extract all relevant information from the wide range of discourse references that are frequently relevant for inference. We identified three general cases, and suggested matching operations to obtain the relevant inferences, formulated as tree transformations. Furthermore, our evidence suggests that for practical reasons, the resolution of discourse references should be tightly integrated into entailment systems instead of treating it as a preprocessing step.

A particularly interesting result concerns the interplay between discourse references and entailment knowledge. While semantic knowledge (e.g., from WordNet or Wikipedia) has been used beneficially for coreference resolution (Soon et al., 2001; Ponzetto and Strube, 2006), reference resolution has, to our knowledge, not yet been employed to validate entailment rules applicability. Our analyses suggest that in the context of deciding textual entailment, reference resolution and entailment knowledge can be seen as complementary ways of achieving the same goal, namely enriching T with additional knowledge to allow the inference of H. Given that both of the technologies are still imperfect, we envisage the way forward as a joint strategy, where reference resolution and entailment rules mutually fill each other's gaps (cf. Example 3).

In sum, our study shows that textual entailment can profit substantially from better discourse handling. The next challenge is to translate the the-

oretical gain into practical benefit. Our analysis demonstrates that improvements are necessary both on the side of discourse reference resolution systems, which need to cover more types of references, as well as a better integration of discourse information in entailment systems, even for those relations which are within the scope of available resolvers.

Acknowledgements

This work was partially supported by the PASCAL-2 Network of Excellence of the European Community FP7-ICT-2007-1-216886.

References

Rod Adams, Gabriel Nicolae, Cristina Nicolae, and Sanda Harabagiu. 2007. Textual entailment through extended lexical overlap and lexico-semantic matching. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

E. Agichtein, W. Askew, and Y. Liu. 2008. Combining lexical, syntactic, and semantic evidence for textual entailment classification. In *Proceedings of TAC*.

Nicholas Asher and Alex Lascarides. 1998. Bridging. *Journal of Semantics*, 15(1):83–113.

Alexandra Balahur, Elena Lloret, Óscar Ferrández, Andrés Montoyo, Manuel Palomar, and Rafael Muñoz. 2008. The DLSIUAES team's participation in the TAC 2008 tracks. In *Proceedings of TAC*.

Roy Bar-Haim, Ido Dagan, Iddo Greental, and Eyal Shnarch. 2007. Semantic inference at the lexical-syntactic level. In *Proceedings of AAAI*.

Roy Bar-Haim, Jonathan Berant, Ido Dagan, Iddo Greental, Shachar Mirkin, and Eyal Shnarch amd Idan Szpektor. 2008. Efficient semantic deduction and approximate matching over compact parse forests. In *Proceedings of TAC*.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009a. The fifth pascal recognizing textual entailment challenge. In *Proceedings of TAC*.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, Medea Lo Leggio, and Bernardo Magnini. 2009b. Considering discourse references in textual entailment annotation. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon (GL2009)*.

Johan Bos. 2005. Recognising textual entailment with logical inference. In *Proceedings of EMNLP*.

Aljoscha Burchardt, Marco Pennacchiotti, Stefan Thater, and Manfred Pinkal. 2009. Assessing the impact of frame semantics on textual entailment. *Journal of Natural Language Engineering*, 15(4):527–550.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of ACL-IJCNLP*.

Nathanael Chambers, Daniel Cer, Trond Grenager, David Hall, Chloe Kiddon, Bill MacCartney, Marie-Catherine de Marneffe, Daniel Ramage, Eric Yeh, and Christopher D. Manning. 2007. Learning alignments and leveraging natural logic. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

Herbert H. Clark. 1975. Bridging. In R. C. Schank and B. L. Nash-Webber, editors, *Theoretical issues in natural language processing*, pages 169–174. Association of Computing Machinery.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Lorand Dali, Delia Rusu, Blaz Fortuna, Dunja Mladenic, and Marko Grobelnik. 2009. Question answering based on semantic graphs. In *Proceedings of the Workshop on Semantic Search (SemSearch 2009)*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2008. The fourth pascal recognizing textual entailment challenge. In *Proceedings of TAC*.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference-6: a brief history. In *Proceedings of the 16th conference on Computational Linguistics*.

Sanda Harabagiu, Andrew Hickl, and Finley Lacatusu. 2007. Satisfying information needs with multi-document summaries. *Information Processing & Management*, 43:1619–1642.

Stefan Harmeling. 2009. Inferring textual entailment with a probabilistically sound calculus. *Journal of Natural Language Engineering*, pages 459–477.

Marti A. Hearst. 1997. Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of HLT-NAACL*.

Jian Huang, Sarah M. Taylor, Jonathan L. Smith, Konstantinos A. Fotiadis, and C. Lee Giles. 2009. Profile based cross-document coreference using kernelized fuzzy relational clustering. In *Proceedings of ACL-IJCNLP*.

Fangtao Li, Yang Tang, Minlie Huang, and Xiaoyan Zhu. 2009. Answering opinion questions with random walks on graphs. In *Proceedings of ACL-IJCNLP*.

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

Dekang Lin. 1993. Principle-based parsing without overgeneration. In *Proceedings of ACL*.

Bill MacCartney, Michel Galley, and Christopher D. Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of EMNLP*.

Katja Markert, Malvina Nissim, and Natalia N. Modjeska. 2003. Using the web for nominal anaphora resolution. In *Proceedings of EACL Workshop on the Computational Treatment of Anaphora*.

Seung-Hoon Na and Hwee Tou Ng. 2009. A 2-poisson model for probabilistic coreference of named entities for improved text retrieval. In *Proceedings of SIGIR*.

Rodney D. Nielsen, Wayne Ward, and James H. Martin. 2009. Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering*, 15(4):479–501.

Massimo Poesio, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004. Learning to resolve bridging references. In *Proceedings of ACL*.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the HLT*.

Long Qiu, Min-Yen Kan, and Tat-Seng Chua. 2004. A public reference implementation of the rap anaphora resolution algorithm. In *Proceedings of LREC*.

Lorenza Romano, Milen Kouylekov, Idan Szpektor, Ido Dagan, and Alberto Lavelli. 2006. Investigating a generic paraphrase-based approach for relation extraction. In *Proceedings of EACL*.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Stephanie Strassel, Mark Przybocki, Kay Peterson, Zhiyi Song, and Kazuaki Maeda. 2008. Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction. In *Proceedings of LREC*.

Jose L. Vicedo and Antonio Ferrndez. 2006. Coreference in Q&A. In Tomek Strzalkowski and Sanda M. Harabagiu, editors, *Advances in Open Domain Question Answering*, pages 71–96. Springer.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Alessandro Moschitti. 2009. A machine learning approach to textual entailment recognition. *Journal of Natural Language Engineering*, 15(4):551–582.

Dmitry Zelenko, Chinatsu Aone, and Jason Tibbetts. 2004. Coreference resolution for information extraction. In *Proceedings of the ACL Workshop on Reference Resolution and its Applications*.