Hunting for Entailing Pairs in the Penn Discourse Treebank

SARA TONELLI[1] *ELENA CABRIO*[2]
(1) Fondazione Bruno Kessler, Trento, Italy
(2) INRIA, Sophia Antipolis, France
satonelli@fbk.eu, elena.cabrio@inria.fr

Abstract

Given the growing amount of resources developed in the NLP community, it is crucial to exploit as much as possible annotated data and tools across different research domains. Past works on discourse analysis have been conducted in parallel with research on semantic inference and, although the two fields of study are intertwined, there have been only few initiatives to put them into relation. Our work addresses the issue of interoperability by investigating the connection between implicit *Restatement* relations in the Penn Discourse Treebank (PDTB) and Textual Entailment. We compare the performance of two TE systems on the *Restatement* pairs and we argue that TE is a subclass of *Restatement* through a manual validation of the pairs. Furthermore, we observe that entailing pairs extracted from the PDTB add interesting and additional levels of complexity to TE, since inference relation relies less on lexical-syntactic variations, and more on reasoning.

Title and Abstract in Italian

A caccia di inferenze semantiche nel Penn Discourse Tree Bank

Data l'ingente quantità di risorse sviluppate in trattamento automatico del linguaggio, l'importanza di sfruttare anche in altri campi di ricerca i dati annotati e gli strumenti implementati è diventata fondamentale. In passato, lavori sull'analisi del discorso sono stati condotti parallelamente alla ricerca sulle inferenze semantiche, ma sebbene i due campi di studio presentino numerosi punti in comune, non ci sono state iniziative per avvicinarli. Questo lavoro affronta la questione dell'interoperabilità investigando le connessioni tra la relazione implicita di *Restatement* nel Penn Discourse Treebank (PDTB) e Textual Entailment (TE) (implicazione semantica). Comparando i risultati ottenuti da due sistemi che riconoscono automaticamente la relazione di implicazione, e dall'annotazione manuale di un sottoinsieme di coppie, mostriamo come il TE sia riconducibile ad una sottocategoria di *Restatement*. Inoltre, osserviamo che le coppie in relazione di implicazione estratte dal PDTB mostrano un livello di complessità superiore rispetto a quelle considerate dai sistemi attuali, in quanto la relazione di inferenza si basa meno su variazioni lessico-sintattiche, e più sul ragionamento.

Keywords: Textual entailment, Penn Discourse Treebank, implicit relations.

Keywords in Italian: Implicazione testuale, Penn Discourse Treebank, relazioni implicite.

Proceedings of COLING 2012: Technical Papers, pages 2653–2668,
COLING 2012, Mumbai, December 2012.

1 Introduction

Given the growing amount of resources and automatic systems developed in the NLP community, it is crucial to guarantee the highest possible compatibility among them, and to exploit as much as possible annotated data and tools across different research domains.[1] Past works on discourse analysis have been conducted in parallel with research on semantic inference (in particular, on textual entailment phenomena, see Sammons et al. (2010), Bentivogli et al. (2010)) and, although the two fields of study seem to be intertwined, no effort has been made into the reuse of annotated data and processing tools across both domains.

With this work, we address the issue of interoperability by investigating the connection between implicit *Restatement* relations in the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) and the Textual Entailment (TE) relation as defined by Dagan et al. (2009). Consider for instance the following sentences extracted from the PDTB and annotated as being in a *Restatement* relation:

(1) *Because hurricanes can change course rapidly, the company sends employees home and shuts down operations in stages.*
The company doesn't wait until the final hours to get ready for hurricanes.

(2) *He's not a reformer – he wants to have the image of a reformer.*
He doesn't want to have the image of the gun man.

Both in Example (1) and (2), a person reading the first sentence would infer that the following is most likely true, which is literally the definition of the textual entailment relation, i.e. a directional relation between a coherent textual fragment (T) and a language expression, which is considered as a hypothesis (H). Entailment holds, i.e. T ⇒ H, if the meaning of H can be inferred from the meaning of T, as interpreted by a typical language user (Dagan et al., 2009). In the examples above, *'The company doesn't wait until the final hours [...]'* and *'He doesn't want to have the image of the gun man'* may be both inferred from the previous sentence by a typical language user, without the need of specific background knowledge.

Since in the PDTB more than 3,000 pairs have been labeled as having an implicit *Restatement* relation, it would be important to assess if they can be used also as training instances for TE systems, and if they represent categories of textual entailment pairs that up to now have not been part of the research agenda of the Recognizing Textual Entailment (RTE) evaluation campaigns[2] due to reasons of convenience for the task definition.

For such challenges, the creation of RTE data sets is a costly and time-consuming activity, requiring a lot of manual work for the creation of the T-H pairs and their annotation (about 1000 training and test instances have been usually provided by the organizers of RTE-1 to RTE-5 challenges). Furthermore, textual pairs extracted from the PDTB would represent real data in a discourse context as opposed to RTE pairs, where T is typically an excerpt extracted from a document (generally newspapers) and H is manually created. On the other hand, if *Restatement* and TE are proved to be equal, RTE technologies could be reused to identify and label this type of implicit relations, which are difficult to detect with existing discourse parsers.

[1] This issue has been recently debated during the Collaboratively Constructed Semantic Resources Workshop's panel at ACL2012, where the importance of the development of functional resources strongly connected with NLP systems was underlined, to allow for resources reusability in different tasks.

[2] `http://aclweb.org/aclwiki/index.php?title=Recognizing_Textual_Entailment`

While the importance of discourse information in TE has already been discussed in relation to anaphora and bridging phenomena (in particular starting from RTE-5, where the T was composed by longer paragraphs, requiring coreference resolution (Mirkin et al., 2010b) (Mirkin et al., 2010a)), little attention has been paid to discourse relations holding between sentences, although this is strictly connected to the problem of coreference (for further discussion on this topic, see Section 3).

In this work, we address the following research questions:

- What are the main differences between *Restatement* and TE relations, given that they are very similar from a theoretical point of view?
- Is it possible to use RTE systems to identify implicit *Restatement* relations, since current approaches have proved to have some limitations and achieve poor performance?
- From a TE perspective, is it possible to use entailing pairs extracted from the PDTB to train or evaluate TE systems?

We believe that addressing these issues is important both from a computational and a theoretical point of view: the findings of this study would be beneficial to system developers as well as to computational linguists interested in discourse and inference phenomena.

The paper is structured as follows: in Section 2 past work related to the identification of implicit relations in the Penn Discourse Treebank is presented and the task of Recognizing Textual Entailment is introduced. In Section 3 the PDTB is described, with a focus on implicit and *Restatement* relations. In Section 4 we present the experimental setting, introducing the TE systems we used, and then we detail both the first and the second experiment we carried out. Finally, we draw some conclusions and discuss future work in Section 5.

2 Related work

A number of approaches have been proposed for annotating *explicit* discourse relations following the PDTB paradigm. While the first attempts were limited to retrieving the heads (usually the main verb) of discourse arguments (Elwell and Baldridge, 2008; Wellner and Pustejovsky, 2007), or to extracting only the sentences containing the arguments (Prasad et al., 2010), more recent works have focused on the identification of the exact arguments spans and on the development of end-to-end discourse parsers (Lin et al., 2010; Ghosh et al., 2011b,a). These works rely on the information conveyed by explicit connectives, which proved quite easy to classify using syntactic information (Pitler et al., 2008; Pitler and Nenkova, 2009).

If the connective is not overtly expressed, however, the task is more challenging and requires different features compared to explicit relations. Experiments by Lin et al. (2009), Pitler et al. (2009) and Lin et al. (2010) showed that, despite the promising results and the progress with respect to their baselines, there is still room for improvement.

As introduced before (Section 1), the notion of textual entailment has been proposed as an applied framework to capture major semantic inference needs across applications in NLP (Dagan and Glickman, 2004), (Dagan et al., 2009). Given a pair of textual fragments, it considers if a competent speaker with basic knowledge of the world would typically infer the second from the first one. To promote the development of general TE recognition engines, designed to provide generic modules across applications, since 2005 the Recognizing Textual Entailment evaluation campaigns[3] have asked participants to develop a system able to detect an inference relation between T-H pairs. In this applied framework, inferences are performed directly over lexical-syntactic representations of the texts. Current systems mainly rely on Machine Learning techniques (typically SVM), logical inference, cross-pair similarity measures between T and H, and word alignment. The definition of TE captures quite broadly the reasoning about language variability needed by different applications for natural language understanding and processing, e.g. information extraction (Romano et al., 2006), text summarization (Barzilay and McKeown, 2005), and reading comprehension systems (Nielsen et al., 2009). Following this rationale, the data sets provided by the challenge organizers are composed of T-H pairs collected from several applicative scenarios (e.g. Question Answering, Information Extraction, Information Retrieval, Summarization), reflecting the way by which the corresponding application could take advantage of automated entailment judgement.[4]

3 *Restatement* relations in the Penn Discourse Treebank

The Penn Discourse Treebank (PDTB) is a resource built on top of the Wall Street Journal (WSJ), in which discourse relations have been manually identified and classified. A *discourse relation* holds between two and only two text spans called *arguments*, that correspond to propositions, events and states.[5]

In the PDTB, relations can be explicitly signaled by a set of lexically defined connectives (e.g. "because", "however", "therefore", etc.). In these cases, the relation is overtly marked, which makes it relatively easy to detect using NLP techniques (Pitler et al., 2008). A relation between two discourse arguments, however, does not necessarily require an explicit connective, because it can be inferred also if a connective expression is missing. These cases are referred to as *implicit relations*, and in the PDTB they are annotated only between adjacent sentences within parahraphs. In case the connective is not overt, PDTB annotators were asked to insert a connective to express the inferred relation.

Examples (3) and (4) represent sentences connected, respectively, by an explicit and an implicit relation. The abstract objects involved in a discourse relation are called `Arg1` and `Arg2` according to syntactic criteria and are reported in italics and in bold respectively.[6]

(3) Explicit: *Use of dispersants was approved* when **a test on the third day showed some positive results.**

(4) Implicit: *The projects already under construction will increase Las Vegas's supply of hotel rooms by 11,795, or nearly 20%, to 75,500.* **By a rule of thumb of 1.5 new jobs for each new hotel room, Clark County will have nearly 18,000 new jobs**.

While in Example (3) the connective "when" explicitly signals a relation holding between `Arg1` and `Arg2`, in (4) no connective was originally expressed. A consequence relation is inferred

[3] `http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool`

[4] Trying to face more real scenarios, in recent editions of the challenge, i.e. RTE-6 and 7, more complexity was added to the traditional main task, asking TE systems to find all the sentences that entail a given H in a set of documents about a topic.

[5] In order to study entailment phenomena between arguments, we make the assumption that arguments correspond to full clauses or sentences. However, in some cases arguments are just textual fragments shorter than clauses.

[6] This notation convention will be applied to all examples reported in this paper, extracted from the PDTB.

between '*the increase in the number of rooms*' and '*the increase in the number of jobs*', though no *explicit* connective expresses this relation.

Each implicit and explicit relation is assigned a sense label based on a three-layered hierarchy of senses. The top-level, or *class level*, includes four major semantic classes, namely TEMPORAL, CONTINGENCY, COMPARISON and EXPANSION. For each class, a more fine-grained classification has been specified at *type* level. For instance, the relation in Example (1) belongs to the CONTINGENCY class and the *Cause* type. A further *subtype* level has been introduced to specify the semantic contribution of each discourse argument.

In this work, we focus on sentence pairs connected by an *implicit* relation and belonging to the EXPANSION class. In particular, we are interested in the relations in the EXPANSION class marked as *Restatement*, because the way in which such relation is defined shows high similarity with the textual entailment relation.

A *Restatement* relation is annotated between two arguments when the semantics of `Arg2` restates the semantics of `Arg1` and it is inferred that the situations described in `Arg1` and `Arg2` hold true at the same time. *Restatement* relations are further specified into three subtypes, namely "specification", "generalization" and "equivalence". The subtype label depends on the ways in which `Arg2` restates `Arg1`: $||\text{Arg1}|| \rightarrow ||\text{Arg2}||$ in the case of generalization, $||\text{Arg1}|| \leftarrow ||\text{Arg2}||$ in the case of specification, and $||\text{Arg1}|| \leftrightarrow ||\text{Arg2}||$ in the case of equivalence, with → indicating logical implication. If more than one subtype interpretation is possible, annotators were allowed to provide a *type* instead of a subtype label, therefore some relations are just classified as *Restatement*.

While intuitively the equivalence relation shares more commonalities with the definition of paraphrase since they both represent bidirectional relations, the specification and generalization types seem to fit well into the definition of textual entailment provided in Section 1 (where the relation of specification has to be considered as a reverse entailment, i.e. the second textual fragment entails the first one).

Let's consider three PDTB sentences annotated as *Restatement*, specifically as implicit specification (5), generalization (6) and equivalence (7):

(5) *She was the child of relative privilege*. **Her mother was a translator; her father was the eternal vice director**.

(6) *Chinese and foreign economists now predict prolonged stagflation: low growth and high inflation.* **The economy is crashing hard**.

(7) *It was like someone had turned a knife in me*. **I was dumbfounded**.

We can represent them in the format of TE pairs, setting as T the first textual fragment and as H the second one (reversing the order for the specification type, as explained before):

(5') T: **Her mother was a translator; her father was the eternal vice director**.
H: *She was the child of relative privilege*.

(6') T: *Chinese and foreign economists now predict prolonged stagflation: low growth and high inflation.*
H: **The economy is crashing hard**.

(7') T: *It was like someone had turned a knife in me*.
H: **I was dumbfounded**.

For all the three pairs, a human reading T would infer that H is most likely true (i.e. they are positive TE pairs).

Leaning on these observations, our intuition is that such pairs automatically extracted from the PDTB could therefore be used to train TE systems, integrating the data sets provided by the RTE challenges organizers (in Section 4 experiments are carried out to prove our intuition). Similarly to RTE pairs, also these sentences are extracted from newspapers. But, while the creation of new T-H pairs requires quite a lot of manual work (for the creation of the H and subsequent annotation), the PDTB is an already available resource. Moreover, in line with the direction of RTE challenges that are now moving toward more real scenarios providing entire documents as T (see RTE-5 to 7), PDTB sentences represent good examples of real data.

Partially guided by reasons of convenience for the task definition, some assumptions have been defined by the organizers of RTE challenges, as for instance the a priori truth of the texts, and the same meaning of entities mentioned in T and H. From a human perspective, the inference required are in general fairly superficial, since generally no long chains of reasoning are involved. Pairs extracted from the PDTB would therefore add interesting and additional levels of complexity to the task, since the relation of inference between T and H relies less on lexical-syntactic variations, and more on reasoning. For instance, Examples (8) and (9) (labeled as *Restatement.equivalence* in the PDTB) can still be considered as positive TE examples, even if they require a lot of background knowledge (e.g. knowledge of idioms and metaphors) for their resolution.

(8) T: **Yet for all his cynicism, he's at heart a closet idealist, a softy with a secret crush on truth, justice and the American Way**.
H: *He's the kind of guy who rescues trampled flags*.

(9) T: *It was like flying without a pilot in the front of the plane*.
H: **It was crazy**.

Even if such level of complexity is still not afforded by current TE systems, the study of these types of arguments could bring new light into textual inference, encouraging the exploration of cases that up to now have not been part of the research agenda.

In the next section we carry out an experimental study *i)* to evaluate the performances of current TE systems on the pairs extracted from the PDTB, and *ii)* to better understand if the theoretical similarity in the definitions of the *Restatement* and the TE relation is actually proved on real data.

4 Experimental setting

In this section, we first introduce the TE systems we used (Section 4.1), and then we present the two sets of experiments we carried out. The first was performed to verify if implicit *Restatement* relations can be detected using current TE systems (Section 4.2). The second was run on a subset of manually re-annotated sentences (Section 4.3), to further verify the relationship between entailment and *Restatement* relations on a controlled data set.

4.1 TE systems description

In order to analyze the correspondence between the *Restatement* and the textual entailment relation, we run different experiments using two off-the-shelf TE systems: VENSES (Delmonte et al., 2009) and EDITS (Kouylekov and Negri, 2010). We choose these two systems because: *i)* they are freely available, *ii)* they obtained similar performances at the last RTE campaigns, and *iii)* they rely on different NLP approaches: VENSES is a rule-based system incorporating and combining different levels of linguistic information, from lexical to semantic knowledge. EDITS, instead, is a supervised TE system implementing a distance-based framework, whose modular architecture combines distance and similarity algorithms.

4.1.1 The VENSES system

VENSES is a rule-based system for recognizing textual entailment based on a linguistic analyzer and an evaluation module.

The first relies on a number of submodules common in Information Extraction systems, i.e. a tokenizer, a multiword and NE recognition module, a PoS tagger based on finite state automata, an in-built syntactic and semantic parser and a pronominal binding module. It also marks and interprets negation, modals and progressive mood.

The evaluation system uses a cost model with rewards/penalties for T-H pairs, where textual entailment is interpreted in terms of semantic similarity: the closest the T-H pairs are in semantic terms, the more probable is their entailment. Rewards in terms of scores are assigned for each 'similar' semantic element.

The system release used in our experiments, which we downloaded at `http://project.cgm.unive.it/venses.html`, is available in two versions: one assigns higher similarity to sentence pairs according to 'shallow' criteria, while the other accounts for 'deep' semantics.

4.1.2 The EDITS system

The EDITS system (Edit Distance Textual Entailment Suite) is an open-source software package for recognizing TE[7] (Kouylekov and Negri, 2010) implementing a distance-based framework which assumes that the probability of an entailment relation between a given T-H pair is inversely proportional to the distance between T and H (i.e. the higher the distance, the lower is the probability of entailment). Within this framework, the system implements different approaches to distance computation, i.e. both edit distance algorithms (that calculate the T-H distance as the cost of insertions, deletions and substitutions that are necessary to transform T into H), and similarity algorithms (e.g. Word Overlap, cosine similarity). Each algorithm returns a normalized distance score between 0 and 1. At a training stage, distance scores calculated over annotated T-H pairs are used to estimate a threshold that best separates positive from negative examples. Such threshold is then used at a test stage to assign a judgment and a confidence score to each test pair.

4.2 Experiment 1: complete data set

Our first experiment is aimed at checking if implicit *Restatement* relations can be detected using existing textual entailment systems. Given that this relation type in the PDTB is defined in

[7] `http://edits.fbk.eu/`

a similar way to textual entailment, we expect TE systems to label sentence pairs having a *Restatement* relation as 'entailing', while sentence pairs connected through another relation type (*Comparison*, *Contingency* or *Temporal*) are likely to be classified as 'not entailing'.

4.2.1 Data set description

We extract all sentence pairs having an implicit *Restatement* relation in the PDTB and we include them in our data set as positive examples. Then, we extract the same number of pairs from the PDTB having an implicit *Comparison*, *Contingency* or *Temporal* relation (the proportion of the three classes reflects their proportion in the PDTB). These pairs are included in the data set as negative examples. The other pairs labeled as *Expansion* but not belonging to the *Restatement* subtype have not been considered in the experiment.

In order to create sentence pairs resembling Text-Hypothesis pairs from RTE challenges, it was necessary in some cases to change the order of the arguments in the positive examples:

- Sentence pairs connected through a *Generalization* label were kept in their original format. Since ||Arg1|| → ||Arg2||, then the sentence corresponding to `Arg1` was considered the Text and `Arg2` the Hypothesis.
- Sentence pairs connected through a *Specification* label were reversed, with `Arg2` being the Text and `Arg1` the Hypothesis. For instance, given the sentences reported in (10) and connected through an implicit *Specification* relation, we build the T-H pair reported in (11).

  (10) *This is an old story.* (`Arg1`). **We're talking about years ago before anyone heard of asbestos having any questionable properties.** (`Arg2`).

  (11) T: We're talking about years ago before anyone heard of asbestos having any questionable properties.
  H: This is an old story.

- In case of sentence pairs connected through an *Equivalence* relation, the longer sentence in the pair was considered as the Text and the shorter one the Hypothesis. This was done in order to resemble as much as possible the T-H pairs in RTE data sets, where the hypothesis is usually shorter than the text. According to the definition of equivalence as ||Arg1|| ⟷ ||Arg2||, however, the entailment relation should hold in both directions.
- For sentence pairs connected through a generic *Restatement* label, we applied the same rule as for the *Equivalence* cases.

Our data set was built to include as positive examples all implicit *Restatement* relations (of any subtype) extracted from the PDTB. In case a relation was annotated with multiple labels, we selected it only if the first option was *Restatement*, otherwise the sentence pair was not included in the data set, neither as positive nor as negative example. Then, the same number of negative examples was collected, having the same proportion of implicit COMPARISON, CONTINGENCY and TEMPORAL relations as in the PDTB. In the end, the data set comprises *6,244 sentence pairs*, equally divided into positive and negative examples.

Since EDITS needs a training set to learn the threshold that best separates positive from negative pairs, the data set was split into a training and a test set, each being 50% of the complete data set with an equal distribution of the sense labels. VENSES does not require supervision, therefore only the test set was used with this system.

4.2.2 Results and discussion

We run VENSES (both deep and shallow versions) and EDITS on the test set. The training set was used by EDITS to learn the threshold, as explained in Section 4.1.2. We apply two basic configurations of EDITS, i.e. Word Overlap and Cosine similarity algorithms on lemmatized texts (stopwords removed).

We compute a simple baseline based on word overlap between the two arguments: we first calculate for each training pair a similarity score using the Text::Similarity::Overlaps library[8] (we use the F1 value, which is a weighted average between the percentage of overlapping words in the text and the percentage of overlapping words in the hypothesis). Then, we train a simple NaiveBayes classifier using only this value as feature (an SVM classifier was trained as well but achieved poorer performance). The model was further used to classify the pairs in the test set.

We report the results in Table 1.

	Baseline	**VENSES** (d.)	**VENSES** (s.)	**EDITS** (wo)	**EDITS** (cos.)
Positive Pairs	*67.53*	33.33	36.14	55.48	51.48
Negative Pairs	*32.07*	62.46	60.16	48.5	49.94
Overall	*49.88*	**47.89**	**48.15**	**49.33**	**50.06**

Table 1: Systems performances on test set (% correctly classified pairs)

VENSES shows a different performance on negative and positive examples both with the shallow and the deep settings. The system has a conservative approach towards entailment, in that it underestimates the positive examples and overestimates the negative ones. Therefore, its performance is better on non-entailing pairs. EDITS strategy, instead, seems to be better balanced. Nevertheless, the systems do not significantly outperform the baseline, which tends to label as 'entailing' also negative pairs.

Given that the performances we obtained on the PDTB data set are below the average system performances in standard RTE tasks (the accuracy of most of the systems ranges between 55% to 65% for the two-way judgement task), these results may depend on two reasons: *i)* either the entailment phenomena underlying the positive examples are more difficult to detect than those in RTE data sets, due to the fact that PDTB pairs are extracted from a real corpus, or *ii)* our hypothesis that the pairs in a *Restatement* relation express also entailment does not hold.

Manually analyzing a set of pairs from the data set for error analysis, we realized that both reasons we hypothesized are (partially) true. As introduced before, TE definition is based on (and assumes) common human understanding of language, as well as common background knowledge. However, the entailment relation is said to hold only if the statement in the text licenses the statement in the hypothesis, meaning that the content of T and common knowledge together should entail H, and not background knowledge alone. For instance, let's consider Example (12).

(12) T: The earlier generation of our crowd bankers had stressed above all probity, tradition,

[8] http://search.cpan.org/t̃pederse/Text-Similarity-0.08/lib/Text/Similarity/Overlaps.pm

continuity and reputation.
H: They were old-fashioned elegant gentlemen.

Even assuming that in H "They" co-refers with the "bankers" in T, according to the definition of TE this is not a positive example, since the amount of background knowledge to be assumed to judge this pair asides from information provided by T. Quite a lot of the pairs tagged as *Restatement* in the PDTB and present in our data set fall into that category, and cannot therefore be considered as positive textual entailment pairs. At the same time, what the literature assumes as background knowledge to be introduced in the inference process is not completely clear (see the debate among Dagan et al. (2006), Manning (2006) and Zaenen et al. (2005)), making the assignment of the entailment judgment to such pairs particularly difficult.

In order to understand and prove if the low performances obtained by the TE systems in our first experiment are due to the presence of *Restatement* sentences that are negative entailment pairs, we conduct a second experiment on a reduced data set.

4.3 Experiment 2: reduced data set

To verify the correctness of our initial hypothesis, i.e. that the sentences annotated as being into a *Restatement* relation express entailment, we run a second experiment focusing on a manually-annotated subset of pairs, as described in the following sections.

4.3.1 Data set description

To investigate to what extent the *Restatement* type is related to entailment, we manually annotated a subset of sentences randomly extracted from the positive examples of Experiment 1 as being *Entailing*, *Non Entailing* or *Entailing with Coreference*. The first two labels were assigned following the RTE annotation guidelines, while the third one was introduced because we observed that in many cases the content of T and H could be put into relation only assuming that coreference was resolved.

We assign a *Coreference* label and not an *Anaphoric* one because we do not limit this analysis to sentences in which some entities are identical, but we also cover pairs in which some information is implicit, and a coreference relation different from identity holds between the two (e.g. bridging). Since the pairs were extracted from adjacent sentences in real documents, this phenomenon is very frequent, because coreference is frequently used as a device to improve textual cohesion. Note that the antecedent is not always in T, as in the following example showing a *Restatement.generalization* relation, in which "She" in T may be resolved through "Marie-Louise (called Marie Latour in the film)" in H:

(13) T: She was untrained and, in one botched job killed a client. Her remorse was shallow and brief. Although she was kind and playful to her children, she was dreadful to her war-damaged husband; she openly brought her lover into their home.
H: Marie-Louise (called Marie Latour in the film) was not a nice person.

In some cases, one of the two sentences is a direct speech restating what was described in the other sentence. Also for these cases, we adopted the *Coreference* label. As a clarification, we report the sentence pair (14): the pronouns "He" and "me" in H should be resolved by "Mr. Sorrell" and "Mr. Roman" in T.

(14) T: But Mr. Roman flatly denied the speculation, saying Mr. Sorrell had tried several times to persuade him to stay, offering various incentives and in one instance sending a note with a case of wine.
H: He asked me not to resign.

We are aware that this kind of sentences would not be considered entailing in standard RTE data sets, but we decided to mark them as *Entailing with Coreference* because direct speech is very frequent in newswire documents, on which the PDTB is based, and we wanted to account also for these cases.

We annotated 160 sentence pairs for each of the four Restatement subtypes (*Restatement.specification*, *Restatement.generalization*, *Restatement.equivalence* and generic *Restatement*), thus collecting 640 annotated pairs. Two annotators were involved in the task. Each annotated independently 240 pairs, while 160 additional pairs were annotated by both, so as to compute inter-annotator agreement.

While the percentage of agreement between the two annotators is 84%, weighted kappa is 0.59. As a rule of thumb, this is a satisfactory agreement, although it reflects the fact that the great majority of assignments are "not entailing", making the probability of chance agreement very high. It is interesting to note that only in one case annotators disagree on whether an entailing relation is also coreferential, meaning that this distinction is well founded and linguistically motivated. In RTE tasks, agreement (Fleiss' Kappa) is usually around 0.98 after reconciliation. Although it is not directly comparable to our agreement (we apply weighted kappa and we do not perform reconciliation), this may reflect the fact that our hypotheses are not manually defined, thus their level of complexity is higher than in standard RTE tasks. As an example, we report in (15) a *Restatement.equivalence* relation on which the annotators disagree. The sentence in H is a sort of lesson that can be drawn from T:

(15) T: The problem is, if people get down in the dumps, they stop selling.
H: Discouragement feeds on itself.

In order to build the final data set, we removed the pairs on which the annotators disagreed (26 in total) and then merged the other annotations. The data set we obtain includes 614 pairs connected through some type of *Restatement* relation and annotated as *Entailing*, *Not Entailing* or *Entailing with Coreference*. Some statistics on the data are reported in Table 2.

Relation Type	**YES**	**YES-COREFERENCE**	**NO**
Specification	7	7	144
Generalization	26	7	123
Equivalence	39	11	93
Restatement	17	16	124
Overall	89	41	484

Table 2: Statistics on manually annotated data in the reduced data set

We observe that most of the pairs are not entailing, and this is generally due to: *i)* the presence of additional information in H not present in T (so the truth of H cannot be verified), as in Example (16), or to *ii)* the presence of not relevant information in H, as in Example 17.

(16) T: Each right entitles the shareholder to buy \$100 face amount of 13.5% bonds due 1993 and warrants to buy 23.5 common shares at 30 cents a share.
H: Under the offer, shareholders will receive one right for each 105 common shares owned.

(17) T: It responds to it.
H: Arbitrage doesn't cause volatility.

These results provide a first answer to the research questions posed in Section 1: despite the definition of the *Restatement* relation in the PDTB, it does not exactly match with the textual entailment relation. Nevertheless, there is an overlap between the two: the annotation suggests that entailing pairs may be a subclass of *Restatement*. As regards the different *Restatement* types, *Specification* is the least related to entailment, with 91% of the pairs annotated as not entailing. This depends on the fact that T and H appear in the original documents in a reversed order, with H typically containing additional information which is not in T. Surprisingly, the relation type showing highest proportion of entailing pairs is *Equivalence* (35% of the examples), and not *Generalization* (21%). In fact, the latter includes many cases in which H is a sort of motto that cannot be inferred directly from T. *Equivalence*, instead, implies less abstraction and is often a reformulation of T at lexical level. Intuitively, pairs into an Equivalence relation are expected to be paraphrases (bidirectional entailment relation), but it is not always the case. For instance, Example (8) can be considered as a positive example of entailment (T⇒ H) once coreference is resolved (both T and H are talking about the same event), but the opposite does not hold, i.e. H does not entail T.

The pairs manually annotated as entailing (130 in total) were used to build a second data set. To balance this new data set with respect to positive and negative pairs in order to run our experiments, we randomly extracted 130 pairs from the negative examples of Experiment 1 (Section 4.2). In the end, we created a data set including 260 pairs, equally divided into positive and negative examples. Compared to the extended data set, all positive pairs in this reduced version have been manually checked, so that they are certain examples of Restatement *and* entailing relation.

4.3.2 Results and discussion

We re-run EDITS and VENSES on the reduced data set to see whether the systems perform better on correct manually annotated entailing pairs. We prepared two versions of the data set: one includes the pairs as they are, while in the other the entailing pairs marked as coreferring were manually resolved, e.g. pronouns were replaced by their extended form, bridging relations were made explicit, and so on.

Since EDITS requires supervision, we split the data set randomly selecting 160 pairs (balanced with respect to positive and negative pairs) to be used for training and 100 for testing. Again, VENSES was run only on the test set, in both the shallow and the deep configuration. We also computed a word-overlap baseline, as in Experiment 1. Results on both versions of the data set (*with* and *without* coreference resolution) are reported in Table 3.

We observe that when running our experiments on the new data set without coreference resolution, both VENSES and EDITS strongly outperform the baseline, while in Experiment 1 there was no significant difference between the three systems. With VENSES, the deep

	Baseline	**VENSES** (d.)	**VENSES** (s.)	**EDITS** (wo)	**EDITS** (cos)
Without coref. res.					
Positive Pairs	*38.00*	40.00	54.00	60.78	38.81
Negative Pairs	*66.00*	82.00	84.00	59.18	69.17
Overall	*52.00*	**61.00**	**69.00**	**60.00**	**59.00**
With coref. res.					
Positive Pairs	*36.00*	56.00	46.00	48.48	41.1
Negative Pairs	*80.00*	82.00	84.00	49.5	66.14
Overall	*58.00*	**69.00**	**65.00**	**49.00**	**57.00**

Table 3: Systems performances on test set (% correctly classified pairs)

version performs better after coreference resolution, but the shallow one achieves a better performance on the original data. Even if EDITS performances are better then the baseline on the original data set, they drop on the reduced data set (in particular, the configuration based on the word overlap algorithm). One of the reasons for that can be seen in the complexity of the pairs extracted from the PDTB, where lexical overlap between T and H is close to 0, and therefore word overlap algorithms fail to correctly detect the positive entailment pairs. VENSES performances show in fact that a linguistically-motivated system including some semantics in the process performs better on such pairs. Moreover, EDITS is negatively influenced by the small size of the data set, since only 160 pairs are used for system training, while for the RTE challenges about 1000 pairs are used for this goal. As short term future goal, we plan to re-run those experiments exploring more customized configurations of EDITS (combining different edit distance algorithms), and including entailment rules to provide it with semantics. In any case, all approaches (including the baseline) improve significantly compared to Experiment 1. On the manually annotated data set the systems performance is comparable to the performance achieved in RTE tasks.

The results of Experiments 1 and 2 allow us to provide some answers and observations concerning the second and the third research questions posed in Section 1. The second question was asking whether it is possible to use RTE systems to identify implicit *Restatement* relations. Current approaches have proved to have some limitations and achieve poor performance, obtaining on average 0.35 F1 on the detection of implicit *Restatement* relations (Lin et al., 2009). Experiment 2 shows that currently available RTE systems outperform such results, and could therefore represent an interesting direction to explore to accomplish the task of identifying implicit *Restatement* relations on a subset of PDTB sentences (i.e. sentences tagged as *Restatement*, where the first argument entails the second one).

The third research question was asking - from a TE perspective - whether it is possible to use entailing pairs extracted from the PDTB to train or evaluate TE systems. As showed in the error analysis following Experiment 1 (Section 4.3.1), less than 1/4 of the *Restatement* pairs are also entailment pairs. At the same time, this subset of *Restatement* and entailment pairs contain interesting and particularly complex pairs, since the relation of inference between T and H relies less on lexical-syntactic variations and more on background knowledge and reasoning. Moreover, such pairs contain entailment phenomena that up to now have not been part of the research proposed by the RTE evaluation campaigns organizers. For instance, in a set of sentences from the PDTB the second argument (i.e. the fragment we consider as H) is a motto, or a metaphor (see Examples (8) and (9)). The presence of such types of arguments, that are easy to understand and solve for humans thanks to their knowledge of the world but almost impossible for machines, could bring new light into textual inference. Indeed, it would encourage the exploration of categories of entailment phenomena that up to now TE systems are not able to face, but that, due to their frequency in real data, cannot be disregarded.

5 Conclusion and future perspectives

In this paper, we provide an analysis of the relation between *Restatement* and textual entailment in the PDTB. Starting from their (similar) theoretical definition, we empirically verified if systems developed for textual entailment recognition can be successfully used to detect implicit *Restatement* relations. Although this first experiment proved that RTE systems are not effective in the task, a manual annotation of the experimental data set showed that part of the *Restatement* examples are also entailing, suggesting that the relation of textual entailment may be a subset of *Restatement* relation. Therefore, RTE techniques could be explored to solve the task on this subset of sentences. Data annotation allowed us to observe also that the *Specification* subtype is the least correlated to entailment cases and that coreference resolution is crucial in identifying entailing sentences in real texts, in line with past research on this topic (Mirkin et al., 2010a,b).

We further showed that the entailing pairs being in a *Restatement* relation represent interesting and particularly complex pairs, because they contain entailment phenomena that are not yet considered in the data provided for RTE evaluation campaigns. Taking into consideration also these cases extracted from real data would bring new light into research on textual inferences, in line with the most recent editions of RTE challenges that are now moving toward more real scenarios. In order to foster this new interesting research direction, the manually tagged pairs used for Experiment 2 will be made available at `http://hlt.fbk.eu/en/people/tonelli/Resources`.

Several research lines have to be considered as future research. As a first step, we plan to improve our experimental evaluation with different respects: *i)* augmenting the size of the reduced data set, in particular annotating more pairs of the PDTB to obtain more TE pairs for RTE systems training and evaluation; *ii)* customizing the EDITS system configuration to increase its performances; *iii)* experimenting with different available RTE systems to compare several approaches to RTE (e.g. logic-based, Machine Learning) on these particularly complex data and *iv)* including other pairs belonging to the EXPANSION class (but not labeled as *Restatement*) in our data set.

Moreover, we would like to verify if the relation between *Restatement* and textual entailment holds also in the other direction, i.e. if positive pairs in RTE data sets would actually be labeled as cases of *Restatement*, and if one of the *Restatement* subtypes can be more frequently associated with positive pairs. Nevertheless, such an annotation could be problematic because *Restatement* is usually defined in the discourse context and taking two sentences in isolation without the surrounding discourse is not likely to lead to meaningful annotations.

Acknowledgments

This work was partially supported by the EC-funded project EXCITEMENT (FP7 ICT-287923). The authors would like to thank Marco Guerini for providing useful feedback on earlier versions of this work. We thank also the PDTB group and the participants to the PDTB Workshop 2012 at IRCS for valuable discussion and insights.

References

Barzilay, R. and McKeown, K. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–327.

Bentivogli, L., Cabrio, E., Dagan, I., Giampiccolo, D., Leggio, M. L., and Magnini, B. (2010). Building Textual Entailment Specialized Data Sets: a Methodology for Isolating Linguistic Phenomena Relevant to Inference. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta. 19-21 May.

Dagan, I., Dolan, B., Magnini, B., and Roth, D. (2009). Recognizing textual entailment: Rationale, evaluation and approaches. *Natural Language Engineering (JNLE)*, 15(Special Issue 04):i–xvii.

Dagan, I. and Glickman, O. (2004). Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability. In *Proceedings of the PASCAL Workshop on Learning Methods for Text Understanding and Mining*, Grenoble, France. 26-29 January.

Dagan, I., Glickman, O., and Magnini, B. (2006). The PASCAL Recognizing Textual Entailment Challenge. In *MLCW 2005, LNAI Volume 3944*. Springer-Verlag.

Delmonte, R., Tonelli, S., and Tripodi, R. (2009). Semantic Processing for Text Entailment with VENSES. In *TAC 2009 Proceedings Papers*. NIST - National Institute of Standards in Technology.

Elwell, R. and Baldridge, J. (2008). Discourse Connective Argument Identification with Connective Specific Rankers. In *Proceedings of ICSC-2008*, Santa Clara, United States.

Ghosh, S., Johansson, R., Riccardi, G., and Tonelli, S. (2011a). Shallow Discourse Parsing with Conditional Random Fields. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, Chiang Mai, Thailand.

Ghosh, S., Tonelli, S., Riccardi, G., and Johansson, R. (2011b). End-to-End Discourse Parser Evaluation. In *Proceedings of the Fifth IEEE International Conference on Semantic Computing (ICSC 2011)*, Palo Alto, United States.

Kouylekov, M. and Negri, M. (2010). An Open-Source Package for Recognizing Textual Entailment. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010) System Demonstrations*, Uppsala, Sweden. 11-16 July.

Lin, Z., Kan, M.-Y., and Ng, H. T. (2009). Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351, Singapore.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2010). A PDTB-Styled End-to-End Discourse Parser. Technical Report TRB8/10, School of Computing, National University of Singapore.

Manning, C. (2006). Local textual inference: it's hard to circumscribe, but you know it when you see it - and NLP needs it. In *Proceedings of the Eighth International Conference on Computational Semantics (IWCS-8)*, Unpublished manuscript. 25 February.

Marcus, M., Santorini, B., and Marcinkiewicz, M. (1993). Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Mirkin, S., Berant, J., Dagan, I., and Shnarch, E. (2010a). Recognising Entailment within Discourse. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 770–778, Beijing, China. Coling 2010 Organizing Committee.

Mirkin, S., Dagan, I., and Pado, S. (2010b). Assessing the role of discourse references in entailment inference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1209–1219. Stroudsburg, PA, USA.

Nielsen, R. D., Ward, W., and Martin, J. H. (2009). Recognizing entailment in intelligent tutoring systems. *The Journal of Natural Language Engineering, (JNLE)*, 15:479–501.

Pitler, E., Louis, A., and Nenkova, A. (2009). Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 683–691, Suntec, Singapore.

Pitler, E. and Nenkova, A. (2009). Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing*.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., and Joshi, A. (2008). Easily Identifiable Discourse Relations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 87–90, Manchester, United Kingdom.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluations (LREC 2008)*, Marrakech, Morocco.

Prasad, R., Joshi, A., and Webber, B. (2010). Exploiting Scope for Shallow Discourse Parsing. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Robaldo, L., Miltsakaki, E., and Bianchini, A. (2010). Corpus-based Semantics of Concession: Where do Expectations Come from? In *Proceedings of LREC*, pages 3593–3600.

Romano, L., Kouylekov, M. O., Szpektor, I., Dagan, I., and Lavelli, A. (2006). Investigating a Generic Paraphrase-Based Approach for Relation Extraction. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy. 3-7 April.

Sammons, M., Vydiswaran, V., and Roth, D. (2010). Ask Not What Textual Entailment Can Do for You... In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, Uppsala, Sweden. 11-16 July.

Wellner, B. and Pustejovsky, J. (2007). Automatically Identifying the Arguments of Discourse Connectives. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 92–101, Prague, Czech Republic.

Zaenen, A., Karttunen, L., and Crouch, R. (2005). Local Textual Inference: Can it be defined or circumscribed? In *Proceedings of the Workshop on the Empirical Modeling of Semantic Equivalence and Entailment*, Ann Arbor, MI. 30 June.