Considering Discourse References in Textual Entailment Annotation

Luisa Bentivogli[1], Ido Dagan[2], Hoa Trang Dang[3],
Danilo Giampiccolo[4], Medea Lo Leggio[4], Bernardo Magnini[1]

[1]FBK-irst
Trento, Italy
{bentivo,magnini}@fbk.eu

[2]Bar-Ilan University
Ramat Gan, Israel
dagan@cs.biu.ac.il

[3]NIST
Gaithersburg, Maryland, USA
hoa.dang@nist.gov

[4]CELCT
Trento, Italy
{giampiccolo,loleggio}@celct.it

Abstract

In the 2009 Recognizing Textual Entailment challenge a Search Pilot task has been introduced, aimed at finding all the sentences in a corpus which entail a set of given hypotheses. The preparation of the data set for this task has provided an opportunity to better understand some phenomena concerning textual entailment recognition in a natural setting. This paper focuses on some problematic issues related to resolving coreferences to entities, space, time and events at the corpus level, as emerged during the annotation of the data set for the textual entailment Search Pilot.

1 Introduction

Recognizing Textual Entailment (RTE) (Dagan and Glickman, 2004; Dagan et al., 2006) is the task of automatically recognizing that the meaning of one text, termed Hypothesis (H), can be inferred by the content of another, termed Text (T). The task definition states that T entails H if, typically, a human reading T would infer that H is most likely true.

RTE has attracted growing interest among NLP researchers, who have been spurred to further investigate the phenomena involved in the challenge. In the annual RTE challenges proposed so far[1] the task has consisted of recognizing, given a set of manually created T-H pairs, whether each T entails the corresponding H. In the last two challenges, the systems have been allowed to make a further distinction and give a three-way judgment, deciding if (i) T entails H; (ii) T contradicts H, or shows it false; (iii) the veracity of H is unknown on the basis of T (Giampiccolo et al., 2008).

In RTE 2009 a Search Pilot task has been set up, consisting of finding all the sentences that entail a given H in a given set of documents about a topic (referred henceforth as the *corpus*). This task is substantially different from the tasks proposed in previous RTE challenges in several ways. First of all, as the entailing sentences to be retrieved belong to a given corpus of documents, the task reflects a natural distribution of entailment. Moreover, in the traditional exercise where isolated T-H pairs are given, both T's and H's are artificially created in such a way that they do not contain references to information outside the T-H pair. In contrast, in the Search task both H and T are to be interpreted in the context of the corpus as they rely on explicit and implicit references to entities, events, dates, places, etc., mentioned elsewhere in the corpus.

The preparation of the RTE Search task data set –both creating the hypotheses and annotating the data set– has been an important occasion to better understand the phenomena involved in the detection of textual entailment in a natural setting, and to begin analyzing some of the problems that can arise when textual entailment is applied to real data in a real context. Although the range of phenomena involved in textual entailment judgment is wide, including ambiguity resolution and discourse phenomena, we focus our investigation on the resolution of coreferences of entities, time, space and events, which have proved to be crucial for textual entailment annotation in the Search scenario. Accordingly, the paper presents an overview of the phenomena concerning reference resolution and some prob-

[1] For details see http://www.nist.gov/tac/2009/RTE/.

lems that were faced in carrying out the annotation of the Search data set.

After a brief description of the task and the data set in Section 2, Section 3 presents the types of knowledge involved in the entailment annotation. Section 4 lists the criteria followed during the creation of the H's, while Section 5 focuses on some issues concerning reference resolution in textual entailment annotation, analyzing some examples taken from the task's Development Set. In Section 6 an analysis of disagreement between annotators is carried out, and Section 7 draws some conclusions and discusses future activities.

2 The Search Task data set

The Textual Entailment Search task is situated in the Summarization application setting where the H's are based on Summary Content Units that have been created from human-authored summaries for a corpus of documents about a common topic (Nenkova et al., 2007). The Entailment Search task is to automatically retrieve all the entailing sentences (T's) in the same corpus, and is evaluated by Precision, Recall, and F-measure. Correctly retrieving a T that entails an H corresponds to correctly extracting a candidate sentence to be included in the summary of the documents. Furthermore, correctly retrieving *all* the entailing sentences for a given H identifies those sentences that contain redundant information and perhaps should not all be included in the summary.

The Entailment Search data set is based on the data created for the Summarization track of the Text Analysis Conference (TAC)[2]. All the examples, issues, and problems reported in this paper refer to the Search Pilot Development Set, which is composed of 10 topics randomly chosen from the 48 topics of the TAC 2008 Update Summarization task.

For each topic, the Search data consist of (i) a set of 10 newswire documents and (ii) between 6 and 10 Hypotheses created from the human-authored multi-document summaries of the set of documents. Since the sentence is the most relevant unit for the Summarization task, all documents have been manually split into sentences, which represent the T's to be judged for entailment.

The Search Development Set contains 80 H's and 2,538 sentences. Each sentence of a topic has been annotated against each H of the same topic, yielding 20,104 sentence annotations, of which 810 are ”entailment” judgments.

3 Knowledge required to perform entailment annotation

The standard definition of textual entailment is based on, and assumes, prior knowledge. As a rule, for textual entailment to hold it is required that T and some assumed knowledge would entail H, but the assumed knowledge alone should not entail H. This means that H may be entailed by incorporating some prior knowledge that would enable its inference from T, but it cannot be entailed by that knowledge alone.

In the traditional RTE task, two types of knowledge are considered necessary to interpret both H and T and make the entailment judgment, namely:

- linguistic knowledge:

H1: Mine accidents cause deaths in China.

T1: So far this week, four mine disasters have claimed the lives of at least 60 workers and left 26 others missing.
Linguistic knowledge: to claim the lives = to cause the death of some people

- common background world knowledge:

H2: The ice is melting in the Arctic.

T2: The scene at the receding edge of the Exit Glacier in Kenai Fjords National Park in Alaska was part festive gathering, part nature tour with an apocalyptic edge.
World knowledge: Alaska is in the Arctic; the edge of the glacier is receding because the ice is melting.

In order to assign the correct entailment judgment in the RTE Search Task, both H and T are to be interpreted in the context of the corpus of 10 documents, as they rely on explicit and implicit references to entities, events, dates, and places pertaining to the topic. Thus, beside linguistic and world knowledge, it is crucial to acquire a dynamic kind of knowledge concerning all explicit and implied references within the sentence, namely:

- corpus knowledge, needed to resolve all the local and cross-document references:

[2] http://www.nist.gov/tac/2009/Summarization/.

H3: 2003 UB313 is bigger than Pluto.

T3: "It's definitely bigger than Pluto", he said of the body made up of ice and rock.
Corpus knowledge: It = the body = 2003 UB313 (a "planet" code name); he = Michael Brown

As H's refer to the whole corpus of 10 documents, it was decided that when judging a sentence for entailment, coreference knowledge available from the *entire* corpus should be taken into consideration, and not just information contained in previous sentences in the same document.

The same criteria specified above for considering prior knowledge in entailment judgment is also adopted for corpus knowledge, i.e. there should be some prior knowledge, possibly including corpus reference knowledge, such that Knowledge and T would entail H, but Knowledge alone should not entail H.

4 The creation of Hypotheses for the Search Task

In the traditional RTE task it is assumed that –in the absence of clear countervailing evidence– mentions of entities, events, places, and dates in H and T corefer. In the Search scenario, where an entire corpus of 10 documents is considered, this simplification is not possible. However, given that H's are manually created, we tried to facilitate the recognition of possible coreferences between H and T by fixing the following criteria:

- H's must be as explicit as possible to reduce ambiguities and facilitate their correct interpretation;
- H's must remain as concise as possible, to maintain linguistic "fluency";
- H's are anchored to the time at which the summaries were written, conventionally fixed at the day after the publication of the last document in the corpus.

Some practical rules based on these criteria were followed in the creation of the Hypotheses. For instance, in mentioning entities, the most complete proper names were used. So, in wording the Hypothesis, "Michael Brown discovered 2003 UB313", the first name and the surname for the scientist and the official scientific denomination for the planet were preferred to other ways of referring to those entities in the corpus.

As far as temporal setting is concerned, some H's contain explicit dates (e.g., "Dennis Rader was arrested on February 25, 2005"). In other cases, the tense of the verb and the implicit time anchor for H disambiguate the temporal context of the event described in H. For example, the Hypothesis "The ice is melting in the Arctic" is presumed to refer to ice melting on 2005/08/15, the date immediately following the last document in the corpus.

Similarly, space specifications were made whenever required for the entailment judgment, especially when ambiguous cases could arise. For example, in the Hypothesis "Mine accidents cause deaths in China", China was explicitly mentioned in order to exclude entailment by sentences mentioning mine accidents in other countries.

Dealing with mentions of events was a little more difficult. Most of the time a short phrase defining the event was used, e.g., "The Kansas Bureau of Investigation collected hundreds of DNA swabs related to the BTK case".

In other cases, the event was defined by the use of the definite article referring to the topic of the corpus, as in the Hypothesis "About 50 people were killed in the attack", where "the attack" implicitly refers to the London bombing event reported in all the documents in the corpus.

5 Coreference resolution in the entailment annotation of sentences

In the following subsections, we will present a number of issues that must be dealt with in order to resolve the references contained in the sentences, and the impact of these phenomena on the entailment annotation. We will focus our attention on the most common and pervasive categories of context-dependent references, i.e. references to entities, events, time, and space.

5.1 Entity coreference resolution

In order to perform entity coreference resolution, it must be taken into account that entities are referred to in a wide variety of ways. As an example, given the Hypothesis "Michael Brown discovered 2003 UB313", the person entity "Michael Brown" is mentioned throughout the corpus in different ways, including variations of proper names such as "Michael Brown", "Michael E. Brown", "Michael A. Brown"[3], "Mike

[3] Note that *Michael A. Brown* is a journalist's mistake.

Brown", "Brown", the pronoun "he", and different types of definite descriptions, such as "the astronomer", and "the Caltech professor".

In the same way, the planet referred to in the Hypothesis is mentioned in various ways in the corpus: "2003 UB313", "UB313", "Xena" (nickname), "it", "the planet", "the new planet", "the possible new planet", "an icy rocky object", "the object", "a lump of rock and ice", "the 10th planet of the solar system". It must be noted that the planet is almost always mentioned with a description, while its proper name, 2003 UB313, appears in only 4 documents out of the 10 in the corpus, showing that it is necessary to resort to cross-document coreference to correctly interpret the sentence and thus judge the entailment.

In some difficult cases the correct identification of the referent becomes crucial to give the correct entailment judgment. Consider the example below:

> *H4:* 2003 UB313 has a moon.
>
> *T4a:* The astronomers who claim to have discovered the 10th planet in the solar system have made another intriguing announcement: it has a moon.
> *Entity coreference:* it = the 10th planet in the solar system = 2003 UB313
> *ENT: YES*
>
> *T4b:* It is different from the previously discovered Kuiper bodies in that it has a moon that circles it every 49 days in a highly elliptical orbit.
> *Entity coreference:* it = EL61 (not 2003 UB313)
> *ENT: NO*

In T4a and T4b the piece of information relevant for entailment is identical *–it has a moon–* and what determines the entailment judgment is the correct reference interpretation of the pronoun "it".

Another interesting phenomenon which arose during the annotation of the corpus sentences was the metonymic use of some expressions in the text, as the following example shows:

> *H5:* The European Union was concerned about freedom of expression in Turkey.
>
> *T5*: Rehn said the new penal code "does not provide sufficient protection for the freedom of expression" and the Turkish government should "close the loopholes in the code."
> *Entity coreference:* Rehn = Olli Rehn = EU Enlargement Commissioner
> *Background/linguistic knowledge:* EU Enlargement Commissioner = European Union (metonymy)
> *ENT*: *YES*

As "Olli Rehn" is considered a metonymic use of "European Union", the sentence T5 is annotated as entailing H5.

5.2 Resolution of time references

Usually, temporal and spatial information in natural texts is given at the beginning of the document to place the story in its space-time coordinates, and then, in subsequent sentences, it is often assumed and not expressed explicitly.

If explicit, time expressions can be absolute or deictic. While absolute expressions are unambiguous, deictic expressions must be normalized, anchoring them to a previous text portion or to the time of utterance, i.e. the date of publication of the article to which the sentence belongs, as the following example shows:

> *H6:* BTK resurfaced in 2004 after a period of silence.
>
> *T6:* The killer has sent several letters to police since resurfacing last March after years of silence.
> *Time Reference*: publication date of the article: 2005/01/06 → last March = March 2004
> *ENT: YES*

As can be seen, the entailment judgment relies on the correct normalization of the deictic temporal expression "last March".

The same anchoring must be performed in the presence of implicit (not expressed) time references.

For non-punctual events, if a precise time range is not given, the reader of a text makes some inferences regarding the typical duration of the event reported, on the basis of his/her prior knowledge. The following example shows how annotators decided about the probable duration of the event mentioned in H, and whether that event is the same as the one reported in the T's:

> *H7*: Ice is melting in the Antarctic.
> [Time anchor: 2005/08/15]

T7a: Last month, scientists again sounded an alarm bell on the effect of global warming on Antarctica, saying that more than 200 coastal glaciers are in retreat because of higher temperatures.
Time Reference: publication date of the article: 2005/06/06 → last month = May 2005
ENT: YES

T7b: Of the 244 marine glaciers that drain inland ice on the Antarctic peninsula, a region previously identified as vulnerable to global warming, 87 percent have fallen back over the last half century, according to research by British experts.
Time Reference: publication date of the article: 2005/06/06 → the last half century = (approximately) second half of 20th century
ENT: YES

T7c: The researchers, reporting in Nature, the British science weekly, say that since the end of the last Ice Age, some 11,000 years ago, the iceshelf had been intact but had slowly thinned, by several dozen metres (several dozen feet).
Time Reference: publication date of the article: 2005/08/03 → 11,000 years ago = 9,000 BC
ENT: NO

The "ice-melting" event in H7 is conceptualized as a long duration event, so the events described in T7a and T7b, while anchored in the previous month or in the previous half century with respect to the publication date of the articles, entail the event described in H7, which is expressed in the present progressive tense and anchored in the conventional date of August 15, 2005. In T7c, instead, the described event was happening 11,000 years ago, and the time reference is perceived as too distant to allow this event to be considered as the same one described in H7.

Another issue concerning time references is specifically related to the fact that the different time anchors of H's and T's have an impact on the interpretation of the verb tenses used. The verb tenses are intrinsically deictic and depend on their anchor time, which for H's is conventionally fixed, while for T's coincides with the publication date of the article.

This anchor mismatch can lead to different situations. If the document is published after the event described in H, it shows the same verb tenses as H, and the entailment judgment is straightforward:

H8: Orhan Pamuk *went* to court on 16 December 2005.

T8: A Turkish court *convened* here Friday to try prominent Turkish author Orhan Pamuk on charges of insulting the Turkish nation, but it was not immediately clear whether the trial would proceed.
Time Reference: publication date of the article: 2005/12/17 → Friday = 16 December 2005
ENT: YES

If the publication of the document is contemporaneous to the event described in H, T will contain verbs in the present tense. Thus it can happen that a sentence containing verbs in the present tense perfectly entails an H containing verbs in the past tense, as in the example below:

H9: The conditions in subway tunnels after the attack *hampered* rescue operations. [time anchor: 2005/07/12]

T9: "It *is* extremely hot and very dusty and it *is* a great challenge for them to continue their work to recover the remaining bodies from the train underground," British Transport Police Deputy Chief Constable Andy Trotter told a news conference.
Time Reference: publication date of the article: 2005/07/09
ENT: YES

Finally, cases where the document is published before the event described in H lead to particularly difficult entailment judgments.

In such cases, the H refers to the event in present or past tense, asserting that the event did happen, whereas T refers to the same event in future tense, describing it as forthcoming. As future tenses involve the issue of modality, the entailment judgment relies essentially on the perception of certainty the annotator has while deciding whether the future event described in T can be considered certain enough to infer the truth of H. See the following example:

H10: Orhan Pamuk *went* to court on 16 December 2005.

T10a: "It is not Orhan Pamuk who *will stand* trial tomorrow, but Turkey."

Time Reference: publication date of the article: 2005/12/15 → tomorrow = 16 December 2005
ENT: YES

T10b: Pamuk, 53, Turkey's best-known novelist**,** *is expected to go* on trial Friday for stating in a Swiss magazine interview what most historians regard as unassailable facts […].
*Time Reference***:** publication date of the article: 2005/12/15 → Friday = 16 December 2005
ENT: NO

Given that in both cases the trial is most likely to happen as scheduled for the following day, the different entailment judgments are due to the interpretation of the linguistic expressions: a verb in simple future tense (*will stand trial*) is perceived as more certain than an epistemic verb (*is expected to go on trial*).

5.3 Resolution of space references

As in the case of time, space references are often implicit at the sentence level. However, our data set contains many H's where spatial information must be recovered in the sentences in order to correctly make the entailment judgment.

Given the two following Hypotheses describing "ice-melting" events in two different locations:

H11a: The ice is melting in the Arctic.

H11b: The ice is melting in the Antarctic.

the assignment of the correct space reference to the following T's becomes crucial to entailment annotation:

T11a: Dressed in tank tops and shorts –beachwear, in fact– on this freakishly warm day in early June, people moved ever closer to the rope line near the glacier as it shied away, practically groaning and melting before their eyes.
*Space Reference***:** the glacier = the Exit Glacier
*Corpus/world knowledge***:** the Exit Glacier is in the Arctic
ENT for H11a: YES
ENT for H11b: NO

T11b: Of the 244 glaciers that drain inland ice and feed these shelves, 87 percent have fallen back since the mid-1950s, according to a British study published in April.
Space Reference (implicit): 244 glaciers and inland refer to the Antarctic.
ENT for H11a: NO
ENT for H11b: YES

5.4 Event coreference resolution

Also when dealing with events, correctly resolving cross-document coreference makes a difference in correctly assessing textual entailment. For instance, consider the following example taken from a topic about the 2005 London bomb attack:

*H12***:** Each bomb used in the attack contained less than 10 pounds of explosives.

*T12***:** Forensic evidence indicates that the bombs each contained less than 10 pounds of high explosives –the Madrid bombs weighed 17 to 22 pounds– which could have kept them small enough to fit in rucksacks, Commissioner Blair said.
Event coreference (implicit): the first mention of bombs refers to the (London) attack.
ENT: YES

In T12 there are two different mentions of bombs and, for each of the two, a different weight is given. In order to assign a correct entailment judgment, it is essential to recognize that only the first refers to the same attack mentioned in H12, while the second refers to another attack, carried out in Madrid.

The example above also presents an interesting lexical issue related to the use of the word "attack". In fact, like the other Hypotheses based on this topic, it refers to a terrorist attack which happened in London and consisted of four distinct attacks. As the term "attack" itself may refer both to a hostile event in its entirety as well as to many sub-events which combined together make up that event, in the T's both the global event and its sub-events are described with the term "attack", either in the singular or in the plural. Since in the H's for this topic only "attack" in the singular is used, searching for entailing sentences may be particularly challenging, because it is necessary to consider sentence by sentence whether "the attack" in the H corefers to the mentions of "attack", both in the singular and in the plural, contained in different T's. Consider the following example:

H13: Bombs were used in the attack.

T13a: Asked whether people should try to forget about Thursday's bomb attack, Livingstone –who often uses public transport– replied: "We carry on our lives.
Event Reference: Thursday's bomb attack = the attack
ENT: YES

*T13b***:** If the terrorists' main aim in London last week was simply to kill people with bombs on public transport, their attacks were a grim success.
Event coreference: attacks = the attack
ENT: YES

T13c: Investigators believe that the materials used to make the rucksack-based bombs were manufactured using smuggled, military-grade explosives, possibly brought in from the Balkans, the paper said.
Event coreference (implicit): bombs refer to the attack.
ENT: YES

As can be seen, H13 is entailed by all the T's presented, even though in T13a the mention of attack is in the singular form, in T13b is in the plural form, and in T13c there is no mention of the attack at all. In each case, only the correct resolution of coreference between the attack/attacks, explicitly or implicitly mentioned in the T's, and "the attack" mentioned in the H allows to determine whether T entails H.

6 The impact of coreference resolution on inter-annotator agreement

In order to assure the creation of a high quality resource, the whole Search Pilot data set was annotated by three assessors. Once the annotation was performed, a reconciliation phase was carried out to eliminate the cases of annotators' mistakes and leave only real disagreements. After the reconciliation phase, 68 cases of disagreement remained, and the inter-annotator agreement calculated using the Kappa statistics (Siegel and Castellan, 1988; Fleiss, 1971) was 0.97. [4]

Disagreements have been classified according to their causes, in order to assess the impact of corpus knowledge. Seven main sources of disagreement were found, namely:

- Meaning of T: different interpretation of the meaning of the Text
- Meaning of H: different interpretation of the meaning of the Hypothesis
- Inference: some annotators make inferences whereas others do not.
- Corpus knowledge: different extent of corpus knowledge used by the annotators
- World knowledge: different amount of common background world knowledge possessed and/or used by the annotators
- Modality: different perception of the certainty of T with respect to the H
- Mixed: this category has been used if more than one factor impacts the different assessments.

Table 1 presents the distribution of disagreements in the corpus:

Cause	Number	%
Inference	19	27,9%
Mixed	15	22,1%
Meaning of T	13	19,1%
Corpus Knowledge	8	11,8%
Meaning of H	6	8,8%
World Knowledge	5	7,4%
Modality	2	2,9%
TOTAL	**68**	100%

Table 1. Classification of disagreements.

Most disagreements are due to the different inferences made by the annotators, while corpus knowledge is the cause of 11.8% of disagreements. In the mixed category, some of the disagreements can also be due to corpus knowledge, so the actual impact of it must be considered higher.

Detailed analysis of the disagreements involving corpus knowledge revealed a phenomenon that is critical for the entailment annotation. These 8 disagreements are not due to a different resolution of references (which is quite clear for humans); instead, the disagreements are due to

[4] Given the high portion of NO judgments in the data set, it is worth mentioning that the percentage of agreement over those annotations where at least one assessor said YES was 92%.

the extent to which corpus knowledge can be used to support a “Yes” judgment for entailment.

As explained in Section 3, H may be entailed by incorporating some prior knowledge, but it should not be entailed by that knowledge alone. So it must be assumed that the incorporated corpus knowledge alone should not entail the H. Nevertheless, since it is difficult to set an a priori boundary to this criterion, annotators in some cases disagree on the extent of corpus knowledge that can be used to assess a YES judgment for a given sentence. Consider the example below:

> *H14*: Sea levels are rising.
>
> *T14*: But what has happened since did.
> *Discourse analysis*: previous sentence: “Because an ice "shelf" already floats on the sea, displacing its weight in water, Larsen B's disintegration - and that of the smaller, nearby Larsen A in 1995 - didn't raise ocean levels.”→ did = raise ocean levels.
> *ENT: ann1: NO, ann2: NO, ann3: YES*

In this case, the core information (“sea level rise”) is recovered from the corpus. However, corpus knowledge alone does not entail the H: it is only the assertion in T, in addition to knowledge, that enables to entail H. Thus, by the strict entailment definition the judgment should be YES, but in practice the annotators did not agree in deciding whether the information necessary for the entailment to hold was all recovered from the corpus or was also contained in the Text.

Such cases suggest the need for further investigation about the limits to which references and other types of knowledge should be extended.

7 Conclusions and future work

In this paper some phenomena concerning the impact of corpus knowledge and reference resolution on the annotation of the Development Set of the 2009 RTE pilot task have been presented. This new exercise introduces an innovative way of performing textual entailment recognition as a search task against a corpus. The data set created represents a useful resource which reflects the natural distribution of entailment in a corpus and presents the problems that can arise while detecting textual entailment in a natural context setting.

We are planning to enrich this resource by annotating all the entailing sentences contained in the Search task data sets with respect to the different types of coreference involved in the entailment, and creating an augmented data set where all the resolved coreferences are made explicit in the sentences.

Moreover, as de Marneffe et al. (2008) have shown in an analysis based on the data used in the previous RTE challenges, contradiction detection plays an important role in text comprehension, and reference resolution is essential to this process. Thus, for future campaigns we are planning to extend the Search task, requiring the retrieval of all the sentences in the corpus that either entail or contradict the H’s. A preliminary analysis of contradiction cases in the Development Set has confirmed that, even though the number of contradicting sentences is very low in the corpus, reference information is fundamental to distinguish between “not entailing” and “contradicting” sentences occurring in a real corpus.

References

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. In *PASCAL Workshop on Learning Methods for Text Understanding and Mining*, Grenoble, France.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In Quiñonero-Candela, J.; Dagan, I.; Magnini, B.; d'Alché-Buc, F. (Eds.), *Machine Learning Challenges,* Lecture Notes in Computer Science, Vol. 3944, Springer.

Marie-Catherine de Marneffe, Anna N. Rafferty and Christopher D. Manning. 2008. Finding Contradictions in Text. In *Proceedings of ACL-08:HLT, Columbus (OH), USA.*

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. In *Psychological Bulletin*, 76(5).

Danilo Giampiccolo, Hoa T. Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio, Bill Dolan. 2009. The Fourth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of TAC 2008*, Gaithersburg (MD), USA.

Ani Nenkova, Rebecca Passonneau, Kathleen McKeown. 2007. The Pyramid Method: Incorporating human content selection variation in summarization evaluation. In *ACM Transactions on Speech and Language Processing,* 4(2).

Sidney Siegel and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences,* McGraw-Hill, New York.