The Seventh PASCAL Recognizing Textual Entailment Challenge

Luisa Bentivogli[1], Peter Clark[2], Ido Dagan[3], Danilo Giampiccolo[4]

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

[2]Vulcan Inc.
Seattle, WA, USA
peterc@vulcan.com

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

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

Abstract

This paper presents the Seventh Recognizing Textual Entailment (RTE-7) challenge. This year's challenge replicated the exercise proposed in RTE-6, consisting of a Main Task, in which Textual Entailment is performed on a real corpus in the Update Summarization scenario; a Main subtask aimed at detecting novel information; and a KBP Validation Task, in which RTE systems had to validate the output of systems participating in the KBP Slot Filling Task. Thirteen teams participated in the Main Task (submitting 33 runs) and 5 in the Novelty Detection Subtask (submitting 13 runs). The KBP Validation Task was undertaken by 2 participants which submitted 5 runs. The ablation test experiment, introduced in RTE-5 to evaluate the impact of knowledge resources used by the systems participating in the Main Task and extended also to tools in RTE-6, was also repeated in RTE-7.

1 Introduction

The Recognizing Textual Entailment (RTE) task consists of developing a system that, given two text fragments, can determine whether the meaning of one text is entailed, i.e. can be inferred, from the other text. Since 2005, the task has been performed in annual RTE challenges which have helped foster the interest of the research community in Textual Entailment. The key to the popularity that the task has gained over the years is that Textual Entailment seems to work as a common framework in which to analyze, compare and evaluate different techniques used in NLP applications to deal with semantic inference.

The first RTE challenge was launched in Europe in 2005 under the aegis of PASCAL, and was followed by two other European rounds. In 2008 RTE became a track at the Text Analysis Conference, joining the efforts with other communities working on NLP applications such as Summarization and Knowledge Base Population (KBP). The collaboration with these other communities was realized by attempting to apply RTE systems to specific application settings. In particular, the RTE-5 Pilot Search Task represented a decisive step forward, as Textual Entailment recognition was (i) performed on a real

text corpus for the first time, and (ii) set up in the Summarization setting, in the attempt to analyze the potential impact of Textual Entailment on a real NLP application.

The RTE-6 challenge tried to capitalize on the advances made in the formulation of the Textual Entailment exercise, and pursued two aims, namely (i) to propose data sets which reflect the natural distribution of entailment in a corpus and present all the typical problems raised by Textual entailment performed in a natural setting; and (ii) to further explore the contribution that RTE engines can provide to Summarization applications. In order to achieve these objectives, some innovations were introduced. First, the traditional Main Task, in which the data sets were composed of isolated, artificially created T(ext) – H(ypothesis) pairs, was replaced by a new Main Task that consisted of recognizing textual entailment within a corpus. The task was situated in the Summarization setting as a close variant of the RTE-5 Pilot Task: given a corpus, a hypothesis H and a set of "candidate" sentences retrieved by the Lucene search engine from that corpus for H, RTE systems are required to identify all the sentences that entail the H among the candidate sentences.

Second, in RTE-6 a Novelty Detection Subtask aimed at specifically addressing the needs of the Summarization Update scenario was also included, where a system had to judge whether the information contained in each hypothesis H is novel with respect to (i.e. not entailed by) the information contained in the corpus. Another major innovation in RTE-6 was represented by the KBP Validation Pilot Task, in which RTE systems had to validate the output of systems participating in the KBP Slot Filling Task (an advanced Information Extraction task). The goal of this was to show the potential utility of RTE systems for Knowledge Base Population and Information Extraction.

As in RTE-5, ablation tests on the knowledge resources and tools used by participating systems were also required, with the aim of studying the relevance of such resources in recognizing Textual Entailment.

In order to ensure the continuity with the previous challenge and allow participants to address the novelties introduced for the first time in RTE-6, the same tasks were used in RTE-7 without significant changes, following the well-established practice of not significantly changing the task every year.

This paper describes the preparation of the data sets for the Main, Novelty Detection and KBP Validation tasks, the metrics used for the evaluation of the systems' submissions, and a preliminary analysis of the results of the challenge. In Section 2 the Main Task is presented, describing the data sets, the evaluation methodology, and an analysis of the results achieved by the participating systems. Section 3 is dedicated to a detailed presentation of the Novelty Detection Subtask, and the KBP Validation Task is described in Section 4. In Section 5 the RTE-7 ablation tests, together with the RTE Knowledge Resources initiative, are presented. Conclusions and perspectives on future work are outlined in Section 6.

2 The RTE-7 Main Task: Recognizing Textual Entailment within a Corpus

Textual Entailment is defined as a directional relationship between two text fragments - T, the entailing text and H, the entailed text - so that *T entails H if, typically, a human reading T would infer that H is most likely true* (Dagan et al., 2006).

This definition of entailment is based on (and assumes) common human understanding of language as well as background knowledge; in fact, for Textual Entailment to hold it is required that *text and knowledge entail H, **but** knowledge alone cannot entail H.* This means that H may be entailed by incorporating some prior knowledge that would enable its inference from T, but it should not be entailed by that knowledge alone. In other words, H is not entailed if H is true regardless of T.

The task of Recognizing Textual Entailment within a corpus, which was introduced as a pilot task in RTE-5 (see Bentivogli et al., 2009b) and became the Main task in 2006, consists of finding all the sentences in a set of documents that entail a given Hypothesis. In such a scenario, both T and H are to be interpreted in the context of the corpus, as they rely on explicit and

implicit references to entities, events, dates, places, situations, etc. pertaining to the topic[1].

As in RTE-6, the RTE-7 Main Task is situated in the Summarization application setting, which means that (i) the RTE corpus is taken from the 2009 and 2010 Summarization Task data set and (ii) the Hs are standalone versions of sentences in that data set, partly selected among the sentences incorporated into some of the automatic summaries created by the systems participating in the Update Summarization Task[2], and partly taken directly from Summarization data set documents.

The goal of the task is to explore the contribution that RTE engines can make to Summarization. In fact, in a general summarization setting, correctly extracting all the sentences entailing a given candidate statement for the summary (similar to Hs in RTE) corresponds to identifying all its mentions in the text, which is useful for assessing the importance of that candidate statement for the summary and, at the same time, detecting those sentences which contain redundant information and should probably not be included in the summary.

The rest of Section 2 describes the Main Task in detail, presenting a description of the task, the resulting data set, the metrics used to evaluate the systems' submissions and the results obtained.

2.1 Task Description

In the RTE-7 Main Task, given a corpus, a hypothesis H, and a set of "candidate" entailing sentences for that H retrieved by Lucene from the corpus, RTE systems are required to identify all the sentences that entail H among the candidate sentences.

The task is not performed on all the sentences in the corpus, but only on a subset of candidates retrieved by Lucene in a preliminary Information Retrieval filtering phase performed by the organizers while building the data set.

For this filtering phase, the retrieval component has to consider (i) each hypothesis as a query and (ii) the corpus sentences as "the documents" to be retrieved. For this purpose, the Apache Lucene[3] text search engine, Version 2.9.1, was used with the following characteristics:

- *StandardAnalyzer* (tokenization, lowercase and stop-word filtering, basic clean-up of words)
- Boolean "OR" query
- default document scoring function.

Regarding the number of sentences to be considered as candidates for entailment, the same criterion as in RTE-6 was followed: only the 100 top-ranked sentences retrieved by Lucene for each H were included in the data set. This choice was based on a study carried out in 2010 showing that with such setting Lucene achieves a recall of about 0.80, providing a good compromise between a sufficient number of entailing sentences and a manageable amount of annotations needed for the creation of a gold standard – even though, in this way, about 20% of entailing sentences, present in the corpus but not retrieved by Lucene, get lost. Unlike in RTE-6, this year the Lucene ranking score for each H was given to participants as supplementary information regarding the preliminary IR phase.

Note that a certain number of Hs have no entailing sentences in the corpus, and also that some documents in the corpus do not contain any entailing sentences.

The example below presents a hypothesis (H) referring to a given topic, and some of the entailing sentences (T) among the larger set of candidate sentences retrieved by Lucene:

H: Lance Armstrong is a Tour de France winner.

T_1: Claims by a French newspaper that seven-time Tour de France winner Lance Armstrong had taken EPO were attacked as unsound and unethical by the director of the Canadian laboratory whose tests saw Olympic drug cheat Ben Johnson hit with a lifetime ban.
(AFP_ENG_20050824.0557 s_id="1")

T_2: L'Equipe on Tuesday carried a front page story headlined "Armstrong's Lie" suggesting the Texan had used the illegal blood booster EPO (erythropoeitin) during his first Tour win in 1999.
(doc="AFP_ENG_20050824.0557" s_id="2")

T_3: The exploits of seven-times Tour de France champion Lance Armstrong, who is alleged to have used

[1] For an analysis of the relevance of discourse phenomena in Textual Entailment see (Bentivogli et al., 2009a).

[2] In the 2009 Summarization Task, the automatic summaries were an assembly of (sometimes modified) selected corpus sentences rather than synthesized sentences.

[3] http://lucene.apache.org/

the banned blood booster EPO (erythropoietin) in 1999, are also down to the use of other banned substances according to one expert. (doc="AFP_ENG_20050831.0529" s_id="1")

T_4: Armstrong, who retired after his seventh yellow jersey victory last month, has always denied ever taking banned substances, and has been on a major defensive since a report by French newspaper L'Equipe last week showed details of doping test results from the Tour de France in 1999. (doc="AFP_ENG_20050831.0529" s_id="3")

T_5: French sports daily L'Equipe reported Tuesday that Lance Armstrong used the performance-enhancing drug EPO to help win his first Tour de France in 1999, a report the seven-time Tour winner vehemently denied (APW_ENG_20050823.0684 s_id=1)

Note that while only the subset of the candidate entailing sentences must be judged for entailment, these sentences are not to be considered as isolated texts. Rather, the entire corpus to which the candidate entailing sentences belong is to be taken into consideration in order to resolve discourse references and appropriately judge the entailment relation. For instance, the second sentence in the example above (T_2) is considered an entailing sentence because from its context it can be seen that "*the Texan"* and *"Tour"* refer respectively to "*Lance Armstrong"* and *"Tour de France"*, mentioned earlier in the discourse.

2.2 Data Set Description

The RTE-7 Main data set is based on the data created for the TAC 2008 and 2009 Update Summarization Task. The TAC 2008 and 2009 SUM Update data consists of a number of topics, each containing two sets of documents, namely (i) Cluster A, made up of the first 10 texts in chronological order (of publication date), and (ii) Cluster B, made up of the last 10 texts.

The RTE-7 data set is composed of 20 topics, 10 used for the Development Set and 10 for the Test Set. For each topic, the RTE-7 Main Task data consist of:

a) A number of Hypotheses (between 20 and 40) referring to the topic. The Hypotheses are standalone sentences taken from the TAC Update Summarization corpus – i.e. both Cluster A and Cluster B documents.

b) A set of 10 documents, corresponding to the Cluster A corpus.

c) For each H, a list of up to 100 candidate entailing sentences (the Ts) from the Cluster A corpus, together with their location in the corpus and Lucene ranking score.

While Ts are naturally occurring sentences in a corpus and are to be taken as they are, the Hs were slightly modified from the originals so as to make them standalone sentences. The procedure applied for the creation of the Hs is described in the following section.

2.3 Creation of the Hypotheses

In the creation of the Hypotheses, two criteria were followed, namely (i) to be as consistent as possible with the Summarization scenario, and cover as much as possible the content of some automatic summaries by systems participating in the Update Summarization Task and (ii) to provide a sufficient number of entailing sentences, in order to respond to the needs of TE systems. To meet these criterea, we adopted a slightly different H creation methodology from that used in RTE-6. In fact, in last year's challenge the goal was that all the content of the automatic summaries of the 10 best scoring systems[4] participating in the TAC 2009 Update Summarization Task was captured by the Hs. However, in the end not all the content was represented, due to practical constraints, namely (i) keeping the number of Hs within the fixed maximum of 30 and (ii) maximizing the number of entailing sentences, which required leaving out some information that made a H too specific.

In order to improve the coverage of the automatic summaries' content, and at the same time keep the number of Hs manageable, in RTE-7 we chose to cover the prominent majority of the content of only the 3 best scoring systems' automatic summaries (instead of 10 as it was in RTE-6). The same procedure defined last year was followed to cover the vast majority of the information contained in the three summaries. First, all the sentences present in the 3 summaries were collected. When a summary sentence contained several pieces of information, it was divided into simpler content units, which were

[4] According to the Pyramid evaluation results for summaries of Cluster B (see Dang and Owczarzak, 2009).

then rephrased as standalone sentences. For example, from the summary sentence (taken from Topic 828 in Test Set) "*Martha Stewart, who is about to get out of prison, seems to have undergone a makeover on the cover of the latest Newsweek",* the following Hs were created:

- H1213: *Martha Stewart is about to get out of Alderson Federal Prison Camp.*
- H1214**:** *Martha Stewart seems to have undergone a makeover on the cover of the latest Newsweek.*
- H1215: *Martha Stewart seems to have undergone a makeover.*
- H1216: *Martha Stewart was on the cover of the latest Newsweek.*

Moreover, although the original sentences were turned into Hs as verbatim as possible, minor syntactic and morpho-syntactic changes were introduced, if necessary, to produce grammatically correct standalone sentences. Similarly, discourse references were also resolved – for instance, in H1213 above the original phrase "*out of prison*" was resolved as "*out of Alderson Federal Prison Camp.",* on the bases of the context from which the summary sentence was taken.

In order to obtain a sufficient number of entailing sentences, as required for the RTE task, an additional number of Hs was created directly from the Cluster A corpus text snippets, even if not present in the automatic summaries (differently from RTE-6 where the additional Hs were taken from Cluster B).

Regarding time anchoring, T and H are naturally anchored to the publication date of the document from which they are taken. This must be taken into account while interpreting T and H verb tenses, since verb tenses are intrinsically deictic and depend on their anchor time (for more detail, see Bentivogli et al., 2009a).

2.4 The Final Data Set

Note that, unlike in RTE-6, the Main Task data set does not contain all the created Hs. Rather, in order to keep the number of Hs manageable, a selection was made, which did not include some of the Hs created from the automatic summaries that had no entailing sentences.

The Development Set is composed of 10 topics, and contains globally 284 Hs, 91 of which were not taken from the automatic summaries but directly from Cluster A sentences. For each H of a topic, all the candidate entailing sentences (100 at most) had to be judged for entailment, yielding 21,420 sentence annotations, of which 1,136 are "entailment" judgments (note that the same sentence can be a candidate for - and entail - more than one H). 110 Hs do not have entailing sentences, while the remaining 174 have at least one entailing sentence.

The Test Set is also composed of 10 topics, and contains globally 269 Hs, 77 of which were not taken from the automatic summaries but directly from Cluster A sentences. There are 22,426 sentence annotations, 1,308 of which are "entailment" judgments. 83 Hs do not have entailing sentences, while the remaining 186 have at least one entailing sentence.

In order to assure the creation of a high quality resource, the whole data set was annotated by three assessors. Once the annotation was performed, a reconciliation phase was carried out to eliminate annotators' mistakes and leave only real disagreements. After the reconciliation phase, the inter-annotator agreement calculated using the Kappa statistics (Siegel and Castellan, 1988; Fleiss, 1971) was 98.35 % for the Development Set and 98.51 % for the Test Set[5].

2.5 Evaluation Measures

The evaluation was carried out in the same way as in the RTE-6 Main Task. System results were compared to a human-annotated gold standard and the metrics used to evaluate system performances were Precision, Recall, and F-measure.

The official metric chosen for ranking systems was micro-averaged F-measure. Additionally, macro-averaged results for topics were made available to participants. As systems were not forced to retrieve at least one entailing sentence for each topic, in order to calculate macro-averaged results it was decided that, if no sentence was returned for a given topic, the Precision for that topic is 0. Also, as many Hs had no entailing sentences, macro-averaged results for hypotheses were not calculated.

[5] It is worth mentioning that the percentage of agreement over those annotations where at least one assessor said YES was 95.51% for the Development Set and 95% for the Test Set.

RUN	Micro-Average			Macro-Average (by TOPIC)		
	Precision	Recall	F-measure	Precision	Recall	F-measure
BIU1	38.97	47.4	42.77	41.3	48.2	44.48
BIU2	41.81	44.11	**42.93**	43.16	45.12	44.12
BIU3	39.26	45.95	42.34	41.00	47.07	43.83
BUPTTeam1	45.02	44.95	**44.99**	47.53	46.41	46.96
BUPTTeam2	48.93	40.37	44.24	52.22	41.88	46.48
BUPTTeam3	51.99	36.93	43.18	56.21	38.63	45.79
CELI1	41.88	46.56	**44.10**	46.63	47.65	47.14
DFKI1	49.4	37.54	42.66	53.98	38.85	45.19
DFKI2	50.77	37.92	**43.41**	56.03	39.5	46.34
DFKI3	53.07	36.31	43.12	58.85	37.63	45.9
FBK_irst1	52.43	32.19	39.89	56.42	33.55	42.08
FBK_irst2	52.33	31.73	39.50	55.46	32.96	41.35
FBK_irst3	46.59	38.07	**41.90**	51.07	39.86	44.78
ICL1	47.88	21.56	**29.73**	49.23	24.59	32.8
IKOMA1	46.96	49.08	**48.00**	48.94	50.22	49.58
IKOMA2	58.48	30.05	39.70	58.87	31.95	41.42
IKOMA3	46.51	49.46	47.94	48.37	50.53	49.43
JU_CSE_TAC1	58.92	19.95	29.81	66.59	20.74	31.63
JU_CSE_TAC2	26.66	35.55	**30.47**	40.63	35.65	37.98
JU_CSE_TAC3	25.16	36.85	29.90	38.99	36.95	37.94
SINAI1	47.08	8.64	14.60	50.15	9.21	15.56
SINAI2	42.99	3.52	6.50	42.95	3.75	6.89
SINAI3	47.3	8.72	**14.72**	50.6	9.27	15.68
SJTU_CIT1	18.52	27.6	22.17	18.35	27.03	21.86
SJTU_CIT2	16.5	38.3	23.07	16.1	37.24	22.48
SJTU_CIT3	17.92	33.33	**23.31**	17.49	32.49	22.74
te_iitb1	20.67	60.24	**30.78**	25.06	63.11	35.87
u_tokyo1	46.49	43.58	44.99	48.44	45.24	46.78
u_tokyo2	47.55	42.35	44.80	48.75	43.61	46.04
u_tokyo3	46.84	43.58	**45.15**	48.63	45.24	46.87
UAIC20111	45.4	18.12	25.90	54.17	19.03	28.17
UAIC20112	30.21	25.84	**27.85**	35.18	27.51	30.88
UAIC20113	18.04	29.66	22.43	23.78	32.29	27.39

Table 1. Main Task results (in bold Best run of each system)

2.6 Submitted Systems and Results

Thirteen teams participated in the Search Task, submitting a total of 33 runs. Table 1 presents the micro- and macro-averaged results of all the submitted runs. Details about Precision, Recall, and F-measure for single topics can be found in the Notebook Appendix. As regards overall results on micro-average, Table 2 shows some F-measure statistics, calculated both (i) over all the submitted runs and (ii) considering only the best run of each participating group.

A first general analysis of the results shows that, unlike in RTE-6, almost two thirds of the systems performed higher in Precision than in Recall. Considering the difference between Precision and Recall within each run, a large variability is noted between the systems, ranging (on micro-averaged results) from (0.07 *BUPTTeam1*) to 39.57 (*te_iitb1)*. Generally a better trade-off between Precision and Recall was seen as compared with the last challenge.

Five Information Retrieval baselines were also calculated. The results are shown in Table 3. The first four baselines were created considering as entailing sentences respectively the top 5, 10, 15, 20 sentences ranked by Lucene. The fifth baseline considered as entailing sentences all the candidate sentences to be judged for entailment in the Main Task, i.e. the top 100 sentences (at most) retrieved by Lucene.

Table 3 shows that Baseline_5 performed best, scoring an F-measure of 37.41, which is 1.46 points above the average, 4.49 points below the median and 10.59 points below the best system's F-measure.

Comparing the results achieved in RTE-6 and in RTE-7, we see that although the best result is essentially the same, an improvement of the overall performances has been registered, as median and average values are higher both considering all submitted runs and only the best runs for each system. Regarding the baseline,

F-measure	All runs	Best runs
Highest	48.00	48.00
Median	39.89	41.90
Average	35.06	35.95
Lowest	6.50	14.72

Table 2. Main Task F-measure statistics

	Precision	Recall	F-measure
Baseline_5	37.00	37.84	37.41
Baseline_10	27.07	55.20	36.33
Baseline_15	21.15	64.65	31.85
Baseline_20	17.71	71.64	28.40
Baseline_100	5.83	100	11.02

Table 3. Baseline results

a similar trend as in RTE-6 is recorded.

The positive overall performance of the systems confirms that RTE techniques could be used, in addition to single IR techniques, to help summarization systems in detecting sentences that imply each other, and thus removing duplicates.

3 Novelty Detection Subtask

The Novelty Detection Subtask consists of judging if the information contained in each H - drawn from the cluster B documents - is novel with respect to the information contained in the set of Cluster A candidate entailing sentences. If for a given H one or more entailing sentences are found, it means that the content of the H is not new. On the contrary, if no entailing sentences are detected, it means that the information contained in the H is novel.

The Novelty Detection Subtask is aimed at specifically addressing the needs of the Summarization Update Task. In this task, systems are required to write a short summary of a set of newswire articles, under the assumption that the user has already read a given set of earlier articles. In such a setting, it is important to distinguish between novel and non-novel information. RTE engines which are able to detect the novelty of Hs - i.e., find Hs which have no entailing Ts - can help Summarization systems filter out non-novel sentences from their summaries. From the systems' point of view, the Novelty Detection Subtask was similar to the Main Task, and did not require any additional annotations. Rather, the novelty detection decision could be derived automatically from the number of entailing sentences found for each H: when no entailing sentences for that H were found among the Cluster A candidate entailing sentences, then the H was judged as novel. In contrast, if one or more entailing sentences were retrieved for a given H, then the H was judged as non-novel. As

in the Main Task, for non-novel Hs all the entailing sentences had to be returned as justification of the judgment. Given this setting, the participants in the Subtask had the opportunity to tune their systems specifically for novelty detection, without having to change their output format.

Nevertheless, the Novelty Detection Task differed from the Main Task because it contained a different set of Hs from that used for the Main Task (see Section 3.1). Moreover, the system outputs were scored differently, using specific scoring metrics designed for assessing novelty detection. In the following, both the data set and the evaluation metrics are described in detail.

3.1 The Data Set

The Novelty Detection data set is similar to the Main task data set except that it contains a different set of Hs and corresponding candidate sentences. In the Main task, Hs are taken both from Cluster B automatic summaries and Cluster A sentences. However, the Hs that are not taken from the automatic summaries are less interesting from a Summarization perspective, because they typically have relatively numerous entailing sentences in the Cluster A corpus and can be more easily recognized as non-novel by the summarization systems. Therefore, the Novelty Detection data contain *all and only* the Hs taken from Cluster B automatic summaries – specifically representing the prominent majority of the content of the automatic summaries of the three best Summarization systems. [6]

The Development Set is composed of 10 topics, and contains globally 254 Hs. Among them, 159 Hs contain novel information (i.e. they have no entailing sentences), whereas 95 Hs do not contain novel information, with a total number of entailing sentences of 576.

The Test Set is composed of 10 topics, and contains globally 302 Hs. Among them, 195 Hs contain novel information (i.e. they have no entailing sentences), whereas 107 Hs do not contain novel information, with a total number of entailing sentences of 779.

The inter-annotator agreement calculated using the Kappa statistics was 98.21% for the Development Set and 98.06% for the Test Set.

3.2 Evaluation Measures

As in the Main Task, the system results were compared to the human-annotated gold standard. Two scores were used to evaluate the system performance on the Novelty Detection Task, namely:

1) The primary score is Precision, Recall and F-measure computed on the binary novel/non-novel decision. The novelty detection decision was derived automatically from the number of justifications provided by the system - i.e. the entailing sentences retrieved for each H - where 0 implies 'novel', 1 or more 'non-novel'.

2) The secondary score measures the quality of the justifications provided for non-novel Hs, that is the set of all the sentences extracted as entailing the Hs. This type of evaluation is the same as the one carried out for the Main Task, and uses the same metrics, i.e. Micro-averaged Precision, Recall and F-measure.

3.3 Submitted Systems and Results

Five teams participated in the Novelty Detection Task, submitting 13 runs. Table 4 presents the results of the Novelty Detection and Justification scores for all the systems participating in the task. More details about Precision, Recall and F-measure for the single topics can be found in the Notebook Appendix.

For overall results on Novelty Detection and Justification scores, Table 5 shows some F-measure statistics, calculated both over all the submitted runs and considering only the best run of each participating group.

Eleven out of 13 submitted runs scored above 80, and considering only the best runs of each system, three out of 5 systems achieved an F-measure above 86.26. Both average and F-measure values - 85.72 and 86.26 respectively - show a good overall performance of the systems. Unlike in RTE-6, Precision was generally higher than Recall, scoring above 90 in 8 cases out of 13, compared to Recall values which remained

[6] A typical kind of information which was not covered by the Hs are the sources of news (e.g., given the sentence "Norwegian Foreign Minister will visit Sri Lanka, the Norwegian Embassy said in a press release.", the piece of information "the Norwegian Embassy said in a press release" is not included in the H set).

RUN	Evaluation - Micro-Average			Justification- Micro-Average		
	Precision	Recall	F-measure	Precision	Recall	F-measure
BIU1	90.74	75.38	**82.35**	31.47	43.26	36.43
BIU2	91.61	72.82	81.14	32.41	41.85	36.53
BIU3	90.12	74.87	81.79	36.34	40.31	**38.22**
CELI1	88.83	85.64	**87.21**	37.92	33.25	**35.43**
DFKI1	92.16	72.31	81.03	37.55	33.5	35.41
DFKI2	93.38	72.31	81.50	38.36	33.63	**35.84**
DFKI3	91.72	73.85	**81.82**	39.42	31.32	34.91
IKOMA1	88.73	92.82	90.73	51.84	27.09	**35.58**
IKOMA2	86.92	95.38	**90.95**	60.61	20.54	30.68
IKOMA3	88.73	92.82	90.73	51.84	27.09	35.58
JU_CSE_TAC1	80.18	93.33	**86.26**	51.19	16.56	25.02
JU_CSE_TAC2	90.6	69.23	78.49	21.94	33.63	**26.56**
JU_CSE_TAC3	90.37	62.56	73.94	20.37	34.27	25.55

Table 4. Novelty Detection task results (in bold the Best run of each system)

below 80, except in 5 cases (see Table 4). Regarding the difference between Precision and Recall within each single run, it varied considerably, ranging from a minimum of 3.19 in *CELI1* to a maximum of 27.81 in *JU_CSE_TAC3.*

A baseline was calculated in which all the Hs are classified as novel. The baseline scored a Precision of 64.57, a Recall of 100, and a corresponding F-measure of 78.47. This baseline, which indicates the proportion of novel Hs in the Test Set, is below the average, and is outdone by the best run results of all systems.

As far as Justification is concerned, unlike in RTE-6 the results did not align with the performances in the Main Task, the best system scoring an F-measure of 38.22. Table 4 shows that 8 out of 13 runs achieved better Precision than Recall, in some cases also significantly, such as in *IKOMA2* where Precision is 40.07 above Recall.

Overall, the results achieved in RTE-7 show a good improvement with respect to RTE-6. In fact, comparing F-measure values to those recorded last year, the best primary score F-measure raised from 82.91 to 90.95; the median from 78.70 to 86.26 and the average from 72.41 to 85.72.

This confirms that Summarization systems could exploit the Textual Entailment techniques for novelty detection when deciding which sentences should be included in the Update summaries.

4 Knowledge Base Population Validation Task

The experiment carried out in the RTE-6 Knowledge Base Population (KBP) Validation Pilot was performed again In RTE-7. The goal of this task, based on the TAC KBP Slot Filling Task (McNamee and Dang, 2009), is to show the potential utility of RTE systems for Knowledge Base Population, similar to the goals in the Summarization setting, thus representing another step towards the creation of a common framework in the field of text understanding.

4.1 Task Description

The KBP Validation Task is situated in the Knowledge Base Population scenario and aims to validate the output of the systems participating in the KBP Slot Filling Task by using Textual Entailment techniques. The idea of using Textual Entailment to validate the output of NLP systems was partly inspired by a similar experi-

NOVELTY DETECTION			JUSTIFICATION (non novel Hs)	
F-measure	All runs	Best runs	All runs	Best runs
Highest	90.95	90.95	38.22	38.22
Median	81.82	86.26	35.43	35.58
Average	83.69	85.72	33.21	34.32
Lowest	73.94	81.82	25.02	26.56

Table 5. Novelty Detection F-measure statistics

ment, namely the Question Answering Validation Task, performed as a part of the CLEF Conferences from 2006 to 2008 (Peñas et al., 2007).

The KBP Slot Filling Task, on which the Validation Task is based, consists of searching a collection of documents and extracting values for a pre-defined set of attributes ("slots") for target entities. In other words, given an entity in a knowledge base and an attribute for that entity, systems must find in a large corpus the correct value(s) for that attribute and return the extracted information together with a corpus document supporting it as a correct slot filler.

The RTE KBP Validation Task is based on the assumption that an extracted slot filler is correct if and only if the supporting document entails a hypothesis summarizing the slot filler. For example, consider the following slot filler and supporting document returned by a KBP system for the "age" attribute for the target entity "Simon Cowell":

KBP System Input
- Target Entity: *"Simon Cowell"*
- Slot: Age
- Document collection

KBP System Output
- Slot Filler: *"*47*"*
- Supp. Doc ID:
APW_ENG_20070315.1712.LDC2009T13

If the slot filler is correct, then the document APW_ENG_20070315.1712.LDC2009T13 must entail one or more of the following Hypotheses, created from the slot filler:

H1: *Simon Cowell is aged 47.*
H3: *Simon Cowell's age is 47.*
H4: *Simon Cowell is age 47.*
H5: *Simon Cowell is 47 years old.*

In other words, the KBP Validation Task consists of determining whether a candidate slot filler is supported in the associated document using entailment techniques.

Each slot filler submitted by a system participating in the KBP Slot Filling Task results in one evaluation item (i.e. a T-H "pair") for the RTE-KBP Validation Task, where T is the source document that was cited as supporting the slot filler, and H is a set of simple, synonymous Hypotheses created from the slot filler.

A distinguishing feature of the KBP Validation Task is that the resulting T-H pairs differ from the traditional pairs because (i) T is an entire document, instead of a single sentence or a paragraph and (ii) H is not a single sentence but a set of roughly synonymous sentences representing different linguistic realizations of the same slot filler.

Another major characteristic of the KBP Validation Task, which distinguishes it from the other RTE challenges proposed so far, is that the RTE data set is created semi-automatically from KBP Slot Filling participants' submissions, and the gold standard annotations are automatically derived from the KBP assessments.

4.2 Data Set Description

The RTE-7 KBP Validation data set was based on the data created for the KBP 2009, 2010 and 2011 Slot Filling Task. More precisely, the Development Set, consisting of the RTE-6 Development and Test sets merged together, was created from the 2009 and 2010 KBP data, whereas the Test Set was created from KBP 2011 data.

The creation of the RTE-7 KBP Validation Task data set was semi-automatic and took as starting points (i) the extracted slot-fillers from multiple systems participating in the KBP *Slot Filling* task and (ii) their assessments[7].

During a first manual phase, before the automatic generation of the Hs for the data set, several "seed" linguistic realizations of templates were created for each target slot, expressing the relationship between the target entity and the extracted slot filler. For example, given the attribute "origin" belonging to a target entity of type "person", the following templates were manually created:

Template 1*:* ***X***'s origins are in ***Y***
Template 2*:* ***X*** comes from ***Y***

[7] As the Slot Filling task can be viewed as a more traditional Information Extraction task, the methodology used for creating the T-H pairs in this Task was the same as that adopted for the manual creation of IE pairs in the Main Task data sets from RTE-1 to RTE-5. In order to create those IE pairs, hypotheses were taken from the relations tested in the ACE tasks, while texts were extracted from the outputs of actual IE systems, which were fed with relevant news articles. Correctly extracted instances were used to generate positive examples, and incorrect instances to generate negative examples.

Template 3*:* ***X*** is from ***Y***
Template 4*:* ***X*** origins are ***Y***
Template 5*:* ***X*** has ***Y*** origins
Template 6*:* ***X*** is of ***Y*** origin

Then, each slot filler submitted by a system participating in the KBP Slot Filling Task became one evaluation item and was used to automatically create an RTE T-H pair. The T corresponded to the corpus document supporting the answer (as identified by the KBP system), while the H was created by instantiating all the templates for the given slot both with the name of the target entity (X) and the slot filler extracted by the system (Y). Providing all the instantiated templates of the corresponding slot for each system answer meant that each T-H pair does not contain only a single H, but rather a set of synonymous Hs. This setting has the property that for each example either all Hs for the slot are entailed or all of them are not.

The procedure adopted to create the Hs implied that some automatically generated Hs could be ungrammatical. While the Hs' templates are predefined, the slot fillers returned by the KBP systems are strings which can be incomplete, include extraneous text, or belong to a POS which is not compatible with that required by a specific H template. For instance, in the example below, given (i) the H templates for the slot "origin", (ii) the target person entity "Simon Cowell" and (iii) a correct slot filler "British", both grammatical and ungrammatical Hs within the same evaluation item were obtained, i.e.:
H1: Simon Cowell 's origins are in British.
H2: Simon Cowell comes from British.
H3: Simon Cowell is from British.
H4: Simon Cowell origins are British.
H5: Simon Cowell has British origins.

These ungrammaticalities were left in the dataset.

The RTE gold standard annotations were automatically derived from the KBP assessments, converting them into Textual Entailment values. The assumption behind this process is that the KBP judgment of whether a given slot filler is correct coincides with the RTE judgment of whether the text entails the template instantiated with the target entity and the automatically extracted slot filler. As the KBP assessments were 4-valued, a mapping was necessary to convert KBP assessments into entailment values: "correct" and "redundant" KBP judgments were mapped into YES entailment; "wrong" judgments were mapped into NO entailment; and, as "inexact" judgments could result both in YES and NO entailment values, RTE pairs involving "inexact" KBP judgments were excluded from the data set.

As in all RTE data sets, temporal issues arise. However, as no temporal qualifications are defined for the KBP slots, differences in verb tense between the Hypothesis and Document Text in the RTE KBP Validation Task had to be ignored. For example, in the KBP Slot Filling Task, "*Tucson, Ariz.*" is considered a correct slot filler for the "residence" attribute of the target entity "Chris Simcox" if the supporting document contained the text "*Chris Simcox lived in Tucson, Ariz., before relocating to Phoenix*"; therefore, in the KBP Validation Task, the Hypothesis "*Chris Simcox lives in Tucson, Ariz.*" must be considered as entailed by the same document.

4.3 Final Data Set

The RTE-7 Development set consisted of a set of T-H pairs from the combined RTE-6 Development and Test sets[8], from which the following pairs have been removed:

- the pairs generated for the location slots "*place of birth*", "*place of death*", "*residence*", and "*headquarters*", which were present only in the RTE-6 Development set and were replace by more specific slots (e.g., "*city of birth*", "*state or province of birth*", and "*country of birth*"...) in the RTE-6 Test set;
- the pairs generated for the slot "*other family*", which was present only in the RTE-6 Development set, and was not included in the Test set as it overgenerated 'YES' entailments with respect to KBP "Correct" judgments;
- pairs where the Ts were speech transcriptions, which were particularly difficult to

[8] The data for the RTE-6 Development and Test sets were created from the KBP slot-filling system output and slot-filler assessments from KBP 2009 and 2010 respectively. For more details, see (Bentivogli et al., 2010)

process as did not contain punctuation and capitalization
- the pairs where the T's are web documents. [9]

Moreover, a number of other pair types were removed following the criteria used in the RTE-6 data set generation.[10]

The final Development Set contained 24,808 T-H pairs, among which 2,231 pairs were positive examples (entailment value "YES"), and 22,577 were negative examples (entailment value "NO").

RUN	GENERIC		
	Precision	**Recall**	**F-measure**
JU_CSE_TAC2	11.79	49.14	**19.02**
CELI3	10.47	29.05	**15.39**
JU_CSE_TAC1	8.01	97.55	14.80
CELI2	8.72	36.74	14.09
CELI1	8.13	43.7	13.71
	TAILORED		
JU_CSE_TAC2	10.97	55.9	18.34
JU_CSE_TAC3	10.97	55.9	18.34
JU_CSE_TAC1	10.8	56.43	18.13

Table 6. KBP Validation results (in bold the Best run of each system)

The KBP Validation Test Set was created from the KBP 2011 assessments. Once unsuitable pairs were removed, the Test Set was created from the 24,140 KBP 2011 Slot Filling Task assessments., obtaing a total of 23,998 T-H pairs.

4.4 Evaluation Metrics

System results were compared to the gold standard created automatically from the KBP assessments of the systems' output. The system performances were measured calculating Micro-Averaged Precision, Recall, and F-measure.

4.5 Submitted Systems and Results

Two different types of submissions were allowed for this task:

- one for *generic* RTE systems, for which no manual effort was allowed to tailor the generic system to the specific slots (beyond fully automatic training on the Development Set);
- the second for *manually tailored* systems, where additional manual effort could be invested to adapt the systems for the specific slots.

Two groups participated in the task, both submitting runs for generic systems and one submitting tailored runs as well. Eight runs were submitted in total – 5 generic and 3 tailored.

Table 6 presents the results, ranked according to F-measure scores. The median F-measure for the all generic runs is 14.80, meanwhile the average value for best runs is 17.20 (15.00 considering all runs); on manually tailored submissions, the average value is 18.27. Overall, the performances were lower than in RTE-6, where the best F-measure score in the generic task was 25.5. Moreover, this year the best result in the tailored task was lower than in the generic task, while in 2010 the manually tailored systems scored higher than the generic ones. All the systems generally had higher Recall scores than Precision, which was quite low both in generic and tailored systems, except in one case. More details about Precision, Recall, and F-measure for each single Slot are given in the Notebook Appendix.

A baseline which classifies all Ts as entailing their corresponding Hs was calculated. The idea behind this baseline is that it reflects the cumulative performance of all KBP 2010 Slot Filling systems, as the RTE-KBP data set includes only Ts which were proposed as implying the corresponding H by at least one KBP system. The baseline, which also indicates the percentage of entailing pairs in the test set, scored a Precision of 6.42, a Recall of 100, and a corresponding F-measure of 12.07. Both participating systems outperformed the baseline, suggesting that a slot filling validation filter using RTE techniques could be useful for KBP systems.

[9] The decision to remove the pairs where the T's are web documents was taken based on the fact that Web documents are on average significantly longer and require much more time to process, representing a major issue for RTE-6 systems participating in the KBP Validation task.

[10] Namely (i) pairs for which the original KBP assessment was "inexact"; (ii) pairs involving KBP system answers of type "NO_RESPONSE"; (iii) duplicate KBP submissions (same answer and document)

The task proved to be particularly challenging for RTE systems, probably due to the fact that the KBP Validation data set was significantly larger than in the other RTE tasks, and most RTE systems are currently not robust enough to process such a large amount of data.

5 System approaches

The twelve systems for which reports have been submitted do not present any significant novelties in addressing the entailment task, proposing strategies already experimented, even though some interesting variations have been introduced.

5.1 Main task

Machine learning is the approach of choice in eight systems out of twelve, integrating a variety of features and techniques. *BIU* proposes a new version of its transformation-based approach using entailment rules and syntactic motivated operations to perform a sequence of inference steps from T to H, which is finally validated by a confidence model. A large number of systems exploit similarity measures and matching algorithms applied at different levels – lexical, syntactic or semantic. *FBK* attempts an approach which moves from token-level to phrase-level overlap, also using paraphrases from parallel data as the main source of lexical knowledge for mapping. *DFKI* approaches lexical similarity by treating T and H as translation of the same source sentence and using the METEOR score to define feature templates to capture similarity between T and H. *IKOMA* system combines entailment scores calculated by lexical matching with machine learning, using a filtering mechanism aimed at discarding T-H pairs which are not entailing, despite high entailment scores based on lexical similarity. *SJTU-CIT* uses machine learning algorithms combined with knowledge drawn from different resources, such as WordNet, VerbOcean and Wikipedia, and features that quantify lexical, syntactic and semantic level matching between T and H. *u_tokyo* uses different WordNet based similarity measure to determine the entailment judgment. A distance-based approach is implemented by EDITS, an open-source RTE package that was exploited by two groups. *CELI* proposes a new version of it, EDIT-GA, extended with genetic algorithms, i.e. a direct stochastic method for global search and optimization that mimics natural evolution. Also *SINAI* uses EDITS, integrating Personalized Page Rank Vectors (PPVs) by means of rules to provide knowledge about the probability of entailment or contradiction between T and H.

Among the systems which do not use machine learning techniques, two systems adopt a rule-based approach, namely *ICL* -which built an inference model based on entailment rules, using also syntactic analysis tools and lexical and semantic resources-, and *UAIC* –which additionally exploits the notion of predication driven entity matching. The remaining systems are based on similarity measures or matching algorithm: *BUPTT* measures the word overlap between T and H; and *JU_CSE_TAC* system is made up by four modules that perform lexical word matching and measure syntactic similarity over chunks and named entities.

5.2 KBP Validation task

Two systems, *JU_CSE_TAC* and *CELI*, attempted also the KBP Validation Task. *JU_CSE_TAC* experimented with both generic and tailored tasks. For the generic task, they use two different methods, (i) one based on entity and verb matching in H and T; (ii) the other using Lucene fed with entities from H as queries to retrieve relevant documents from the corpus and basing the entailing judgment on whether or not the corresponding T is found in the top retrieved documents. In the tailored task, validation rules extracted from the development data for each attribute are applied. *CELI* first performs a preliminary filtering phase, where sentences in T are selected as entailing candidates only if they contain at least one lexical matching with H; then EDITS-GA trained over development data for optimal configuration is used to assess entailment between each H and the set of entailing candidate sentences selected from the corresponding T.

6 RTE-7 Ablation Tests and RTE Knowledge Resources initiative

Also in RTE-7, ablation tests, introduced first in RTE-5 to perform an evaluation of the impact of

knowledge resources (and from RTE-6 also of tools), were required of participants in the Main task. The idea is that ablating the resources and tools used by a system allows those resources' and tools' contribution to the system's performance to be evaluated. The kind of ablation tests required in the RTE-7 Main Task consists of removing one module at a time from a system, and re-running the system on the test set with the other modules. By comparing these results to those achieved by the complete system, the practical contribution of the individual component can be assessed.

As in previous challenges, the participants responded well to the initiative. In fact, out of 13 participants in the Main task, 10 submitted ablation tests, while three could not carry out any tests because the architecture of the system did not allow the removal of any components. In total, 31 ablations tests were performed and submitted. Despite these guidelines, 7 submitted ablation tests did not specifically ablate knowledge resources or tools, but a variety of other system components, such as entailment algorithms, empirically estimated thresholds, and other statistical features. In two cases, a combination of different components was removed from the system instead of a single one.

Results for all the submitted ablation tests are in the Notebook Appendix. Table 7 gives summary information about the 21 ablation tests complying with our requirements. For knowledge resources, 16 ablation tests were carried out on a total of 7 different resources. For tools, 5 ablation tests were performed to evaluate 2 types of tools. For each ablated component, the number of ablation tests submitted is shown, together with the number of runs showing a negative or positive impact of the resource/tool on the system performance.

Note that while the data provided by the ablation tests can provide an indication of the actual contribution of a component to the performance of a specific system, determining the general impact of a knowledge resource or a tool is not straightforward, since the same resource can be used in different ways by different systems, and thus the results of different ablation tests on the same resource are not fully comparable. For instance, Table 7 shows that a common resource such as WordNet had a small positive impact in most cases - and sometimes it even slightly worsened the system's performance - except in the case of *U_CSE_TAC1_abl1*, in which the removal of WordNet caused a decrease of 9.81 points in the F-measure value. A similar behaviour can be also noticed regarding Wikipedia, another commonly used resource. Detailed data on the ablation test results are provided in the Appendix and can be used for further analysis. In fact, since a better knowledge of how resources and tools can be best utilized within Textual Entailment may be relevant for the research, a study of the ablation test results may help not only to identify which resources were more useful within each system but, considering the architecture of the systems that performed well, also to learn how to effectively use such resources.

Another initiative to promote the study of the impact of knowledge resources on Textual Entailment consists in making available the results of all the ablation tests carried out so far on the RTE Knowledge Resources web page where a list of the "standard" knowledge resources (currently 36 publicly available and 15 non-publicly) used to design RTE systems in the challenges held so far is also found. [11]

Participants are encouraged to help keep the page up-to-date, sharing their own knowledge resources and tools, not only to contribute to the research on the impact of knowledge resources on RTE, but also to have the opportunity to further test and leverage such resources and tools.

7 Conclusions and Future Work

After the major innovations introduced in the RTE-6 challenge - which marked the transition from the traditional Main Task proposed in the first five RTE challenges to a new Main Task in which textual entailment recognition was performed on a real text corpus - RTE-7 followed the well-established practice of not significantly changing the tasks every year, in order to ensure continuity with the previous challenge and allow participants to address the novelties introduced for the first time in RTE-6. So this year the RTE-6 tasks were repeated, introducing only

[11] http://www.aclweb.org/aclwiki/index.php?title=RTE_Knowledge_Resources.

	Ablated resource # Ablation tests		Impact on systems					
			Positive			Negative		
			# Tests	F1 Range	F1 Average	# Tests	F1 Range	F1 Average
Knowledge Resources	WordNet	**8**	**5**	0.64 9.81	2.62	**3**	-0.05 -0.14	-0.10
	VerbOcean	**1**	**1**	5.93	5.93	**-**	-	-
	Wikipedia	**3**	**2**	1.56 8.89	2.25	**1**	-2.64	-2.64
	DIRECT	**1**	**1**	0.94	-	**-**	-	-
	Paraphrase table	**1**	**-**	-	-	**1**	-1.43	-1.43
	CatVar	**1**	**1**	0.84	0.84	**-**	-	-
	Acronym Lists	**1**	**-**	-	-	**1**	-0.16	-0.16
Tools	Coreference Resolver	**1**	**1**	0.69	0.69	**-**	-	-
	Named Entities Recognition	**4**	**2**	2.08 7.97	5.03	**2**	-0.89 -8.29	-4.59

Table 7. Ablated knowledge resources

minor changes.

The Main Task was mainly aimed at further testing textual entailment performed in a corpus, while the Novelty Detection Task was primarily dedicated to explore how TE systems can help Summarization systems to filter out non-novel sentences from their summaries. Similarly, the KBP Validation Task's goal was to show the potential utility of RTE systems for Knowledge Base Population and Information Extraction.

The results in the Main task largely reflected those achieved in RTE-6, but an improvement of the overall performances was recorded. In fact, if on the one hand the best F-measure score was practical identical to that achieved last year, on the other hand the median and average F-measure values were higher than in RTE-6.

The Novelty Detection Subtask confirmed the success of the previous exercise, recording higher best, median and average scores, which further showed that RTE systems perform well in detecting novelty, and could be useful for Summarization systems.

As regards the KBP Validation Task, it demonstrated once again to be the most complex of the tasks proposed, and only two participants took part in the exercise. Even though the reasons which make this task so arduous have not been properly investigated yet, the impression is that a major discouraging factor lies in the difficulty that current RTE systems have in processing large amounts of data.

Overall, the RTE-7 tasks confirmed that entailment systems may play an important and effective role within semantic applications. In fact, the results obtained so far represent an encouragement to further promote the advancement of the state of the art in textual entailment recognition, and to continue the effort to demonstrate its applicability and usefulness in other real-life application scenarios, as it has already done in the fields of automatic summarization and knowledge base population.

Acknowledgments

We would like to acknowledge other people involved in the RTE-7 challenge: Alessandro Marchetti, Giovanni Moretti from CELCT, and Karolina Owczarzak from NIST.

This work is supported in part by the Pascal-2 Network of Excellence, ICT-216886-NOE.

References

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini and Idan Szpektor. 2006. The Second PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment*, Venice, Italy.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, Medea Lo Leggio and Bernardo Magnini. 2009a. Considering Discourse References in Textual Entailment Annotation. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon (GL 2009)*, Pisa, Italy.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, Bernardo Magnini. 2009b. The Fifth PASCAL Recognizing Textual Entailment Challenge. In *TAC 2009 Workshop Notebook*, Gaithersburg, Maryland, USA.

Luisa Bentivogli, Peter Clark, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo. 2010. The Sixth PASCAL Recognizing Textual Entailment Challenge. In *TAC 2010 Workshop Notebook*, Gaithersburg, Maryland, USA.

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.

Hoa Trang Dang and Karolina Owczarzak. 2009. Overview of the TAC 2009 Summarization Track. In *TAC 2009 Workshop Notebook*, Gaithersburg, Maryland, USA.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, Bill Dolan. 2007. The Third PASCAL Recognizing Textual Entailment Challenge, In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic.

Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio. 2008. *The Fourth PASCAL Recognizing Textual Entailment Challenge. In TAC 2008 Proceedings*. http://www.nist.gov/tac/publications/2008/papers.html

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

Paul McNamee, Hoa T. Dang, (2009). Overview of the TAC 2009 Knowledge Base Population Track. In *Proceedings of the TAC Workshop*, Gaithersburg, MD, USA.

Anselmo Peñas, Alvaro Rodrigo, V. Sama, Felicia Verdejo, (2007). Testing the Reasoning for Question Answering Validation. In *Journal of Logic and Computation* http://logcom.oxfordjournals.org/cgi/reprint/exm072

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