The Fifth PASCAL Recognizing Textual Entailment Challenge

Luisa Bentivogli[1], Ido Dagan[2], Hoa Trang Dang[3], Danilo Giampiccolo[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@celct.it

Abstract

This paper presents the Fifth Recognizing Textual Entailment Challenge (RTE-5). Following the positive experience of the last campaign, RTE-5 has been proposed for the second time as a track at the Text Analysis Conference (TAC). The structure of the RTE-5 Main Task remained unchanged, offering both the traditional two-way task and the three-way task introduced in the previous campaign. Moreover, a pilot Search Task was set up, consisting of finding all the sentences in a set of documents that entail a given hypothesis. 21 teams participated in the campaign, among which 20 in the Main Task (for a total of 54 runs) and 8 in the Pilot Task (for a total of 20 runs). Another important innovation introduced in this campaign was mandatory ablation tests that participants had to perform for all major knowledge resources employed by their systems.

1 Introduction

Textual Entailment Recognition is a generic task that captures major semantic inference needs across many Natural Language Processing (NLP) applications, such as Question Answering, Information Retrieval, Information Extraction, and multi-document Summarization, providing a common solution for modeling language variability. 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. Proposed for the first time in 2005, RTE has enjoyed a constantly growing popularity in the NLP community, as it 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, a common issue shared by many NLP applications.

After three successful campaigns held in Europe, characterized by a steady increase in number of participants and results achieved[1], the fourth RTE challenge in 2008 marked a major change with respect to the organization of the campaign, as it was proposed for the first time as a track at the Text Analysis Conference[2] (TAC). The challenge, jointly organized by CELCT and NIST, introduced for the first time a three-way judgment (already proposed as a pilot task in 2007) in the Main task. The final outcome of RTE-4 confirmed the successful trend, and a total of 26 participants took part in the challenge.

In 2009 the fifth round of the RTE challenges presented a mixture of innovation and continuity

[1] For more information see previous RTE challenges' overviews.

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

with the previous competitions. Besides the traditional Main Task, a Pilot Search Task was also proposed, consisting of finding all the sentences in a set of documents that entail a given hypothesis. The two tasks were aimed on the one hand at allowing new and old participants to test their systems against the classic RTE task setting, and on the other at keeping the interest of the research community high by introducing a more realistic scenario, where textual entailment recognition is performed on a real text corpus. Furthermore, ablation tests on the knowledge resources used by systems were introduced as a mandatory requirement for all the participants in the RTE-5 Main Task, with the aim of studying the relevance of such resources in recognizing textual entailment.

This paper describes the preparation of the data sets of both Main and Search tasks, the metrics used for the evaluation of the systems submissions, and an analysis of the results. 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. The Knowledge Resource Pool and the ablations tests are also described. Section 3 presents an overview of the first five RTE challenges, describing how the data sets and the system performances evolved from RTE-1 to RTE-5. Section 4 is dedicated to a detailed presentation of the Pilot Search Task, describing the preparation of the data sets, the metrics used for the evaluation and providing an analysis of the system results. In Section 5 conclusions and perspectives on future work are outlined.

2 The RTE-5 Main Task

The Textual Entailment task requires systems to decide, given a set of text pairs called *T(ext)* and *H(ypothesis)*, whether T entails H or not. Textual entailment is defined as a directional relation between two text fragments – T, the entailing text and H, the entailed text – so that a human being, with common understanding of language and common background knowledge, can infer that *H* is most likely true on the basis of the content of *T*.

The RTE-5 Main Task was kept very similar to that proposed in RTE-4, in order to facilitate the comparison between the performances of systems which had participated in the previous campaign and encourage new participants to take part in an exercise not too different from last year's task. Nevertheless, some changes were introduced in order to move towards a more realistic exercise, stimulating researchers who had already participated in other RTE campaigns to further test their systems against more challenging data sets.

First of all, while the length of the Hs was the same as in the past data sets (around 8 words), in the RTE-5 data set Ts were longer, up to 100 words, whereas in RTE-4 the average length was about 40 words. This length was meant to represent the average portion of the source document that a reader would naturally select, such as a paragraph or a group of related sentences. On the other hand, longer texts introduced in the exercise discourse phenomena, such as coreference, which were not present in the previous data sets. Moreover, texts, taken from a variety of freely available sources to avoid copyright problems, were not edited from their source documents. In this way, systems were asked to handle real text that may include typographical errors and ungrammatical sentences.

For the rest, the basic structure of the challenge remained unchanged[3]. Like in the previous RTE-4 campaign, both the classic two-way task and the three-way task were offered. In the traditional two-way task the pairs where T entails H are marked as ENTAILMENT, and those where the entailment does not hold are marked as NO ENTAILMENT. The three-way task requires to further distinguish, in case there is no entailment between T and H, whether the truth of H is contradicted by T, or remains unknown on the basis of the information contained in T. In other words, the systems participating in the three-way task have to decide whether:

- T entails H (ENTAILMENT judgment)
- T contradicts H (CONTRADICTION judgment)
- The truth of H cannot be determined on the basis of T (UNKNOWN judgment)

[3] For more details on the creation and the characteristic of the datasets, see the RTE-4 overview (Giampiccolo et. al, 2008).

TASK	TEXT	HYPOTHESIS	ENTAILMENT

QA	The Grapes of Wrath, published exactly 70 years ago, can be seen as a prophetic novel, rooted in the tragedies of the Great Depression, but speaking directly to the harsh realities of 2009, writes Steinbeck scholar Robert DeMott. Steinbeck's epic novel, which traces harrowing exodus of Tom Joad and his family from blighted Oklahoma (where they are evicted from their farm), across the rugged American south-west via Highway 66, and on to what they mistakenly hope will be a more promising future in California, is considered by many readers to be the quintessential Depression-era story, and an ironic reversal of the rags-to-riches tale favoured by many optimistic Americans.	"The Grapes of Wrath" was written by Steinbeck.	ENTAILMENT
IE	Henan province has registered seven dead children and 4,761 HFMD cases. Shandong has reported five children dead from HFMD and 3,280 cases to deal with. HFMD can start from a variety of viruses of which Enterovirus 71 (EV-71) is the most common, followed by the Coxsackie A virus (Cox A16). There is an Incubation period from time of contact to appearance of symptoms between three to seven days.	Shandong is not far from Henan province.	UNKNOWN
IR	An appeals court in Eastern France has confirmed the Swedish car manufacturer Volvo is guilty over the deaths of two schoolchildren aged nine and ten and the serious injury of a third after a brakes failure caused an accident in 1999. The Volvo 850 TDI was being driven by a local teacher when it struck the children, who had been on their way to school. Driver Catherine Kohtz later asserted that the brake pedal had become stiff and the brakes themselves unresponsive as she traveled along the steep road.	Volvo is a car manufacturer from Finland.	CONTRADICTION

Table 1. Examples of T-H pairs from the RTE-5 data set

The settings from which the pairs were manually created by human annotators were Information Extraction (IE), Information Retrieval (IR), and Question Answering (QA). Summarization (SUM) was not considered in this year's Main Task, as the Pilot Search data sets were entirely based on the Summarization setting. Table 1 presents some examples of T-H pairs taken from the RTE-5 data set.

The RTE-5 data set consisted of 1,200 T-H pairs - 400 for each setting - equally divided into a Development Set and a Test Set. The distribution according to the 3 way annotation, both in the individual settings and in the overall data set, was 50% ENTAILMENT, 35% UNKNOWN, 15% CONTRADICTION.

As in the previous challenges, the overall process of data set creation requires the generation of large amounts of T-H pairs, which are subsequently filtered to retain only those (i) featuring full agreement among three annotators in terms of the assigned entailment judgment, and (ii) compliant with the RTE guidelines for the creation of entailment pairs. The effort required to create the pairs varies a lot depending on the application scenario (being the QA pairs the most difficult to create and the IR pairs the easiest ones), and the type of entailment pair to be created (entailment, unknown, contradiction). On average, six pairs per hour are created and annotated for the first time by an expert annotator. The subsequent entailment annotation of the existing pairs is much less time-consuming, as forty pairs per hour can be annotated.

As regards the RTE-5 data set, around 25% of the pairs originally created were discarded due to disagreement, and another 20% because they were unsuitable according to the guidelines (e.g. Ts too short or too long, ENTAILMENT pairs with the elements relevant to the entailment judgment repeated verbatim, or UNKNOWN pairs with T and H completely unrelated).

2.1 Evaluation Measures

The evaluation of all runs submitted was automatic, the judgments returned by the system being compared to the Gold Standard compiled by the human assessors.

The main evaluation measure was accuracy, i.e., the fraction of correct answers. For the two-way task, a judgment of "NO ENTAILMENT" in a submitted run was considered to match either

"CONTRADICTION" or "UNKNOWN" in the Gold Standard.

As a second measure, an Average Precision score was computed for systems that provided as output a confidence-ranked list of all test examples. Average Precision is a common evaluation measure for system rankings, and is computed as the average of the system's precision values at all points in the ranked list in which recall increases, that is at all points in the ranked list for which the gold standard annotation is ENTAILMENT. In other words, this measure evaluates the ability of systems to rank all the T-H pairs in the test set according to their entailment confidence (in decreasing order from the most certain entailment to the least certain). More formally, it can be written as follows:

$$\frac{1}{R}\sum_{i=1}^{n}\frac{E(i)\times \# EntailmentUpToPair(i)}{i}$$

where *n* is the number of the pairs in the test set, R is the total number of ENTAILMENT pairs in the Gold Standard, E(i) is 1 if the i-th pair is marked as ENTAILMENT in the Gold Standard and 0 otherwise, and *i* ranges over the pairs, ordered by their ranking.

In practice, the more confident the system was that T entailed H, the higher the ranking of the pair was. A perfect ranking would have placed all the positive pairs (for which the entailment holds) before all the negative ones, yielding an average precision value of 1.

As average precision is relevant only for a binary annotation, in the case of three-way judgment submissions the pairs tagged as CONTRADICTION and UNKNOWN were conflated and retagged as NO ENTAILMENT.

2.2 Submitted Systems and Results

Twenty of the 21 RTE-5 teams participated in the Main Task (1 team participated only in the Search Pilot Task), slightly less than in RTE-4, when the participants were 26.

Participants were allowed to submit runs to one or both of the tasks (2-way and 3-way judgment). Runs submitted to the 3-way task were automatically converted to 2-way runs (where CONTRADICTION and UNKNOWN judgments were conflated to NO ENTAILMENT) and scored for 2-way accuracy. However, participants in the 3-way task were also allowed to submit a separate set of runs for the 2-way task, which need not be derived from any of their 3-way runs. This allowed researchers to pursue different optimization strategies for the two tasks.

In the end, 7 participants took part only in the 3-way task; 10 only in the 2-way task, and 3 in both - which means that almost half of the participants chose to test their systems against the three-way judgment task. The total number of submitted run was 54, among which 24 were evaluated for the 3-way task.

Table 3 presents the results obtained by each participant, both in the 3-way and in the 2-way task. As regards overall results, Table 2 shows some accuracy statistics calculated (i) over all the submitted runs and (ii) considering only the best run of each participating group.

	3-way Task		**2-way Task**	
	All runs	**Best runs**	**All runs**	**Best runs**
Highest	68.33	68.33	73.5	73.5
Lowest	43.83	46.83	50.00	50.00
Median	52.00	55.83	61.08	61.5
Average	52.91	56.1	60.36	61.52

Table 2. Main Task accuracy statistics

In the 3-way task, the best accuracy on best runs was 68.33, only slightly less than last year, when it was 68.50. In contrast, average accuracy, always on best runs, was significantly higher than last year (56.1 compared to 52.59 in RTE-4), indicating that the 3-way task is still difficult but improvement is feasible. A similar trend was recorded in the 2-way task, which as usual proved to be easier than the 3-way. Although best accuracy on best runs was one point lower than last year (73.5 compared to 74.6 in RTE-4), average accuracy was 2 point higher (61.52 compared to 59.41 in RTE-4). It must be noted that a comparison is not really possible as the data sets of RTE-4 and RTE-5, although similar in the creation and structure, were actually different. However, given the increased difficulty of the RTE-5 data set, the results can be considered encouraging.

RUN	3-WAY Accuracy	2-WAY Accuracy	Average Precision	RUN	3-WAY Accuracy	2-WAY Accuracy	Average Precision
AUEBNLP1		**61.00**	55.65	DLSIUAES3-*3way*	47.17	61.50	
AUEBNLP2		60.17	54.78	FBKirst1		**60.17**	59.54
AUEBNLP3		59.83	54.16	FBKirst2		56.33	55.90
AUEBNLP1-*3way*	57.00	61.33	53.15	FBKirst3		56.83	57.33
AUEBNLP2-*3way*	**57.5**	**61.5**	53.53	JU_CSE_TAC1		**58.17**	
AUEBNLP3-*3way*	57.17	61.17	53.01	JU_CSE_TAC2		58.17	55.08
BIU1		63.00		PeMoZa1		64.17	64.26
BIU2		**63.83**		PeMoZa2		**66.17**	65.99
Boeing1		**60.00**		PeMoZa3		61.83	62.37
Boeing2		59.33		QUANTA1		**67.00**	70.11
Boeing1-*3way*	43.83	55.33		QUANTA2		66.33	67.55
Boeing2-*3way*	46.33	56.67		rhodes1-*3way*	**57.00**	**61.00**	
Boeing3-*3way*	**54.67**	**61.5**		Sagan1-*3way*	**52.17**	**55.17**	
clr091		**53.17**		Sagan2-*3way*	52.00	54.5	
cswhu1-*3way*	**52.17**	**63.33**		Sagan3-*3way*	51.83	54.83	
cswhu2-*3way*	52.00	63.33		Siel_091-*3way*	46.00	**60.83**	
DFKI1-*3way*	50.67	62.5		Siel_092-*3way*	46.00	60.83	
DFKI2-*3way*	**63.67**	66.83		Siel_093-*3way*	**46.83**	60.67	
DFKI3-*3way*	63.50	**68.5**		ssl1-*3way*	**48.67**	**56.00**	
DirRelCond1		**61.5**		ssl2-*3way*	44.33	52.33	
DirRelCond2		59.67		UAIC20091-*3way*	**68.33**	**73.5**	
DirRelCond3		59.67		UB.dmirg1		**50.00**	
DLSIUAES1		62.83		UB.dmirg2		50.00	
DLSIUAES2		**63.17**		UB.dmirg3		50.00	
DLSIUAES3		62.00		UI_ccg1		**64.33**	
DLSIUAES1-*3way*	**60.00**	62.00		VensesTeam1		**61.50**	64.45
DLSIUAES2-*3way*	51.00	**62.83**		VensesTeam2		61.50	

Table 3. Main Task results (in bold Best run of each system)

As a final remark, it is worth mentioning that the two top-scoring systems for the 2-way task were actually 3-way runs evaluated for the 2-way task.

As far as per-task results are concerned, Table 4 shows the average accuracy calculated over all the submitted runs. As can be seen, the trend registered in RTE-4 was confirmed, in which IE was found to be the most difficult application scenario, and IR the easiest one.

Task	3-way Task	2-way Task
	Average Accuracy	Average Accuracy
IE	47.25	53.31
QA	51.15	57.45
IR	60.33	70.32

Table 4. Average accuracy by task

2.3 RTE Resource Pool and Ablation Tests

Since many participants at the RTE-4 workshop expressed a widespread interest in the various knowledge resources used by RTE systems, two initiatives were proposed in the RTE-5 campaign aimed at studying the relevance of knowledge resources in recognizing textual entailment. Firstly, a section specifically dedicated to knowledge resources was added to the RTE Resource Pool[4], listing the "standard" knowledge resources that have been selected and exploited in the design of RTE systems during the challenges held so far, together with the links to the locations where they are made available. Furthermore, a shortlist of the "top" resources was also published, as well as some results of the data analyses which were conducted on the resources presented in the page.

Secondly, in order to evaluate the contribution of each single resource to the systems' performances, ablation tests were introduced as a requirement for systems participating in RTE-5 Main Task. An ablation test consists of removing one module at a time from a system, and re-running the system on the test set with the other modules, except the one tested. The idea is that comparing the results achieved in the ablation tests to those obtained by the systems as a whole will allow assessing the contribution given by each single resource. RTE participants were required to run ablation tests on all the knowledge resources used by their systems, and submit the results together with the system runs.

The initiative was successful and had a very positive response, as out of 20 participants in the Main Task only one did not submit any ablation tests (because no knowledge resources were used). In total, 82 ablations tests were performed and submitted. In order for an ablation test to be suitable to our purposes, it had to be carried out (i) removing only knowledge resources and (ii) removing one resource at a time. As a matter of fact, 28 submitted ablation tests did not specifically address knowledge resources but a variety of other system components, such as pre-processing modules, entailment algorithms, empirically estimated thresholds and other statistical features. In other 16 ablation tests, a combination of different resources/components was removed from the system instead of one single resource. Results for all the submitted ablation tests can be found in the TAC Proceedings[5], while Table 5 gives further information about the 38 ablation tests conformant to our requirements. For each ablated knowledge resource, the number of ablation tests submitted is given, together with the number of runs showing a negative, positive and null impact of the resource on the system performance.

Table 5 shows that all the most common knowledge resources used by TE engines were tested by a large number of systems. However, determining the actual impact of these knowledge resources is not straightforward as it happens that systems make different uses of the same resources and hence the results of the ablation tests are not really comparable. As can be seen in the table, resources have a positive impact on some systems and a negative or null impact on other systems. The most evident example is given by WordNet, which among all the evaluated resources turns out to have both the highest positive impact (4% accuracy on two different systems - *Boeing3* and *UI_ccg1*) and highest negative impact (2% accuracy improvement when removed from the *AUEBNLP1* system).

Ablated resource	**Ablation tests**	**Impact on systems**		
		Positive	**Null**	**Negative**
WordNet	19	9	3	7
VerbOcean	6	2	3	1
Wikipedia	4	3	0	1
FrameNet	3	1	1	1
DIRT	3	2	0	1
RTE dataset	1	1	0	0
PropBank	1	1	0	0

Table 5. Ablated knowledge resources

[4] http://www.aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool

[5] http://www.nist.gov/tac/publications/2009/results.html

3 From RTE-1 to RTE-5: an Overview of the First Five RTE Challenges

From its beginning in 2005, the task of Recognizing Textual Entailment has evolved significantly, although its basic structure has been maintained in the years. In the first three challenges the task consisted of assigning a two-way entailment judgment (YES/NO) to a set of T-H pairs. In RTE-4 and RTE-5 an additional 3-way judgment task was proposed together with the original one. In this task, in case of no entailment between T and H, systems have to specify whether T contradicts H (CONTRADICTION judgment), or the truth of H cannot be determined on the basis of T (UNKNOWN judgment).

3.1 The Data Sets

In all the editions of the Challenge, the T-H pairs were created by expert annotators from a number of NLP application settings. In the first campaign the applications considered were Information Retrieval (IR), Comparable Documents (CD), Reading Comprehension (RC), Question Answering (QA), Information Extraction (IE), Machine Translation (MT), and Paraphrase Acquisition (PA). In the following three campaigns they were limited to IE, IR, QA, and Summarization (SUM). In RTE-5 only IE, IR, and QA were considered, and SUM was chosen as the setting for a separate Pilot Search task.

Table 6 shows how the composition of the data sets evolved over the years, in terms of number of pairs, T and H length, and word overlap between T and H. As far as the length of T and H is concerned, while Hs length remained constant over the years, the length of Ts substantially increased, passing from an average of 24.78 words in the RTE-1 Development set to around 100 words in the RTE-5 data sets. This gradual change to longer texts allowed for the introduction of discourse phenomena in the data set, which represented a first step towards the more realistic scenario proposed in the RTE-5 Search Pilot Task, where Textual Entailment was performed against a real corpus.

Table 6 also shows data about the average word overlap between T and H, which is calculated counting all the words shared by T and H, and normalizing the results by the length of H. Overlap rates are grouped on the basis of the entailment judgment (YES/NO) assigned to the pairs. In general, it can be seen that positive examples (entailment=YES) show a higher word overlap with respect to the negative ones. Moreover, it is interesting to analyze the difference in word overlap between positive and negative pairs. This difference steadily increased over the years, reaching its highest value in the RTE-3 data sets, where the average overlap for positive pairs amounts to 71% whereas for negative pairs it amounts to 54%. This suggests that, for systems taking word overlap into account, the RTE-3 data set is potentially easier to process. RTE-4 and RTE-5 data sets are different from the previous ones, due to the introduction of the three-way classification of the pairs. If we consider

Challenge	Data Set	Pairs	H length (# words)	T length (# words)	T/H OVERLAP (%)		
					YES	NO ENTAILMENT	
						Unknown	Contradiction
RTE-1	DEV	567	10.08	24.78	69.25	62.94	
	TEST	800	10.8	26.04	68.64	64.12	
RTE-2	DEV	800	9.65	27.15	69.1	58.16	
	TEST	800	8.39	28.37	70.63	63.32	
RTE-3	DEV	800	8.46	34.98	72.18	53.24	
	TEST	800	7.87	30.06	69.62	55.54	
RTE-4	TEST	1,000	7.7	40.15	68.95	57.36	67.97
RTE-5	DEV	600	7.79	99.49	77.71	61.95	77.06
	TEST	600	7.92	99.41	77.14	62.28	78.93

Table 6. RTE-1 to RTE-5 data sets

the class of NO-ENTAILMENT pairs, on the one hand we see a large difference in word overlap between UNKNOWN and ENTAILMENT pairs (similar to that present in the RTE-3 data set); on the other hand, CONTRADITION pairs present a high word overlap, very similar to that of ENTAILMENT pairs. This makes the RTE-4 and RTE-5 particularly challenging, as a part of the negative pairs are not distinguishable from the positive pairs by simply considering the word overlap feature.

3.2 Two-way Task Results

As far as the analysis of the results is concerned, it must be noticed that only the 2-way task, consisting of assigning a YES/NO entailment judgment, has been proposed in all the five challenges held so far, allowing for a more comprehensive comparison.

In order to better analyze the systems' performances and to assess the difficulty of the different data sets, we exploit the results of one of the eight word overlap baselines proposed in (Mehdad and Magnini, 2009). These baselines are created according to different criteria followed to calculate the overlap between T and H, namely (i) inclusion vs. exclusion of stop words, (ii) use of lemmas vs. tokens, (iii) overlap count normalization vs. no normalization.

For our purposes we present baseline number 8, where stop words are excluded from the overlap count, and neither lemmatization of the words in T and H, nor normalization of the overlap count is performed. Table 7 shows the results of such baseline for all the RTE challenges, together with some accuracy statistics regarding the systems' performances. It can be noticed that, while the baseline results present relatively small differences in four of the challenges (with scores ranging from 54.4% to 57.5%), the result obtained on the RTE-3 data set is considerably higher (reaching 62.4%) and outperforms both the median system and the average accuracy.

The results of all the runs submitted by participants to the five challenges are plotted in Figure 1.

Challenge	Baseline	Median (Best runs)	Average (Best runs)
RTE-1	55.37	56.20	56.45
RTE-2	54.4	59.00	59.87
RTE-3	62.4	61.75	61.97
RTE-4	56.6	58.30	59.41
RTE-5	57.5	61.50	61.52

Table 7. Baseline-8 and system results

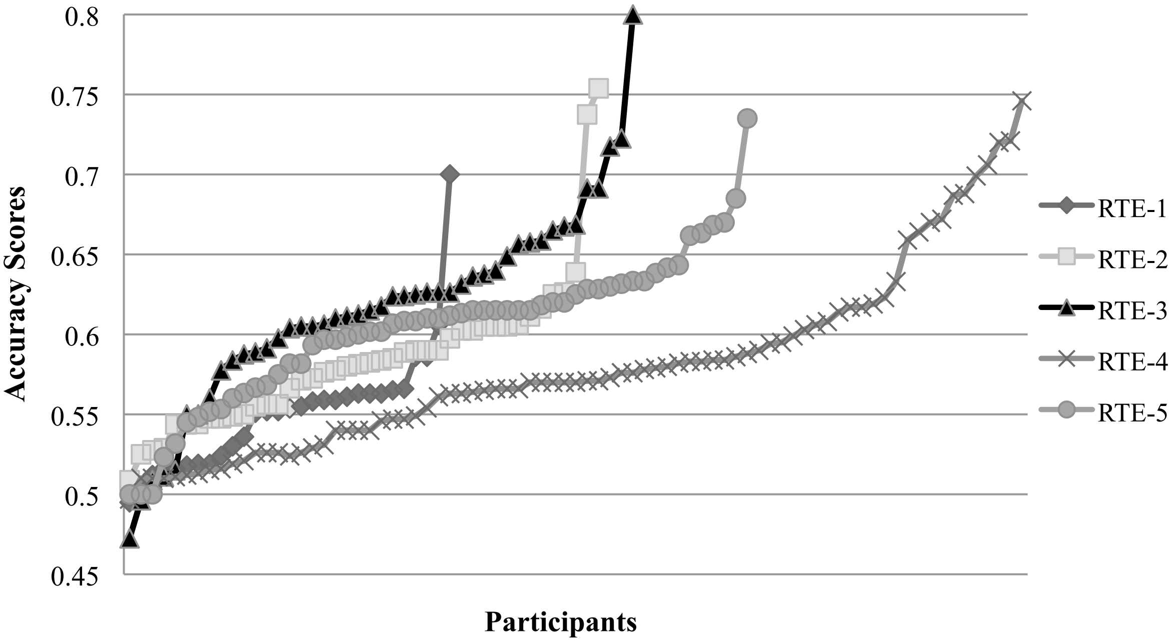

Figure 1. 2-way task results from RTE-1 to RTE-5

As a first remark, it can be said that while the lowest results in the different challenges are distributed over a quite narrow span (ranging from 47.25% in RTE-3 to 50.88% in RTE-2), the interpretation of the trend of the highest scores is more difficult. After a neat improvement of the best results in RTE-2 with respect to RTE-1 (with best accuracy increasing from 70% to 75.38%), a new sensible improvement was registered in RTE-3, reaching an accuracy of 80%. The positive trend anyway was inverted in RTE-4 and RTE-5, in which the best accuracy was 74.6% and 73.5% respectively.

In general, we can see from the curves in Figure 1 that the overall system performances increased from RTE-1 to RTE-3, then they dropped in RTE-4 and increased again in RTE-5, but without reaching the performances shown in RTE-3. This trend is in line with the baselines results, and with the data sets characteristics discussed in Section 3.1. In fact, both the baseline and the systems achieve the highest scores on the RTE-3 data set, which features the highest difference in word overlap between positive and negative pairs, suggesting that the RTE-3 data set was actually the easiest. The results' drop in RTE-4 can be explained by (i) the introduction of the three-way judgment which changed the characteristics of the data set, and (ii) the fact that a development set was not released. It is finally worth noticing that despite the greater difficulty of the RTE-5 data set (not in terms of word overlap, but due to the increased length of Ts), systems showed a slight improvement with respect to the previous year, suggesting an overall advancement of the state of the art.

3.3 Three-way Task Results

As regards the three-way task, Figure 2 compares the results obtained in the two challenges in which it was proposed (RTE-4 and RTE-5). Table 8 shows more analytically that, although the best accuracy scores achieved in the two challenges are very similar (68.50% in RTE-4 and 68.33% in RTE-5), the overall trend was positive and a general improvement can be observed, as both lowest and average accuracy raised (respectively improving from 30.70% to 43.83%, and from 50.65% to 52.91%).

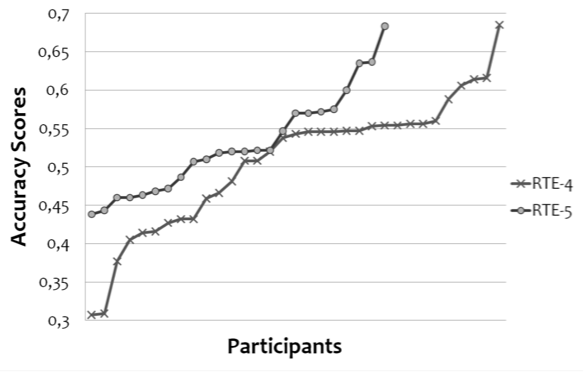

Figure 2. 3-way task results RTE-1 to RTE-5:

Results	RTE-4 Accuracy (%)	RTE-5 Accuracy (%)
Highest	68.50	68.33
Lowest	30.70	43.83
Median	54.30	52.00
Average	50.65	52.91

Table 8. 3-way statistics in RTE-4 and RTE-5

4 The RTE-5 Search Pilot Task

During the RTE-4 workshop at TAC 2008 the need to move towards more realistic scenarios was stressed once more, both by organizers and participants. The earlier RTE campaigns had proposed only test sets in which the pairs had been artificially adapted in order to facilitate the study of the different entailment phenomena, and to allow participants to get acquainted with the textual entailment task; however, the progress in RTE research now allows to make a step forward and start to test the RTE systems against real data.

In order to meet this demand, a Textual Entailment Search Pilot task was set up with a twofold objective. On the one hand, the task was aimed at producing a data set reflecting the natural distribution of entailment in a corpus and presenting all the problems that can arise while detecting textual entailment in a natural setting. On the other hand, the goal was to analyze the potential impact of textual entailment recognition on a real NLP application, namely the Summarization task as proposed by the Summarization community in the TAC 2008 workshop at NIST.

4.1 Task Description

The Search Pilot Task consists of finding all the sentences that entail a given H in a given set of documents about a topic (referred henceforth as the *corpus*).

The Textual Entailment Search Task is situated in the Summarization application setting where (i) the Hs are based on Summary Content Units(SCUs)[6] that have been created from human-authored summaries for a corpus of documents about a common topic, and (ii) the entailing sentences (Ts) are to be retrieved in the same corpus for which the summaries were made. Correctly extracting all sentences entailing a given candidate statement for the summary (similar to our hypotheses) corresponds to identifying all its mentions in the text. The information identified this way, such as the positions and overall frequency of these mentions, is often useful to assess the importance of that candidate statement for the summary. 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.

An example taken from the Search data set is presented below.

Topic *A820* - Russian mini-submarine accident

(i) Hs created starting from the existing SCUs, representing typical content for a summary:

H1:The AS-28 mini-submarine was trapped underwater
H2: Seven submariners were onboard the AS-28
H3: The AS-28 accident happened in eastern Russia
H4:The AS-28 got entangled in fishing nets
H5:The AS-28 crew was rescued in satisfactory conditions.

(ii) Some of the *A820* corpus sentences which entail H2:

H2 Seven submariners were onboard the AS-28

T The Russian military was racing against time early Friday to rescue a small submarine that had become trapped on the seabed with seven crew aboard, the Ria-Novosti news agency reported. *(doc_id="AFP_20050804.0725" s_id="1")*

T All seven aboard the AS-28 mini-submarine appeared to be in satisfactory condition, naval spokesman Capt. Igor Dygalo said. *(doc_id="AFP_20050807.0129" s_id="2")*

T It was carrying six sailors and a representative of the company that manufactured it. *(doc_id="APW_ENG_20050807.0129" s_id="8")*

T The seven men on board were said to have as little as 24 hours of air *(doc_id="NYT_ENG_20050805.0181" s_id="2")*

T There are seven crew members aboard the vessel, stranded on the ocean floor at a depth of around 190 meters (623 feet) in a bay off the coast of the Kamchatka peninsula in Russia's Far East region. *(doc_id="AFP_ENG_20050805.0571" s_id="6")*

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, as can be seen from the above example, a major difference with respect to the Main Task is that in the framework of the traditional exercise, where isolated T-H pairs are given, both Ts and Hs are artificially created in such a way that they do not contain references to information outside the T-H pair, and hence the context necessary to judge the entailment relation is given by T. Only language and world knowledge are considered necessary to interpret both T and H and make the entailment judgment.

In contrast, in the Search Task 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, etc., pertaining to the topic and mentioned elsewhere in the corpus.

Thus, besides linguistic and world knowledge, it is crucial to acquire a dynamic kind of knowledge concerning all explicit and implied references within the sentence, i.e. corpus knowledge, which is needed to resolve all the local and cross-document coreferences. As Hs refer to a whole corpus, 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[7].

[6] SCUs are sub-sentential content units, not bigger than a clause, taken from a set of manually-made summaries. SCUs are used in the evaluation of Summarization tasks. See (Nenkova et al., 2007) and (Dang and Owczarzak, 2008)

[7] For a detailed discussion on the different types of discourse references see (Bentivogli et al., 2009).

4.2 Data Set Description

The Search data set is based on the data created for the TAC 2008 and 2009 Update Summarization task[8]. More precisely, the Development Set topics were randomly chosen from those of the 2008 exercise, whereas the Test Set topics were randomly taken from the TAC 2009 SUM Update data. In this way, the evaluation of both the summarization and the textual entailment systems was carried out contemporaneously and on the same data, making comparisons easier and allowing a more precise analysis of the possible impact of textual entailment recognition on the summarization task.

For each topic, the Search data consist of (i) a corpus 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 Ts to be judged for entailment. While Ts are naturally occurring sentences in a corpus and are to be taken as they are, the Hs have to be manually created. The procedure applied for the creation of the Hs is described in the following section.

4.2.1 The Creation of Hypotheses

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 Hs are manually created, we fixed the following criteria aimed at facilitating the recognition of possible coreferences between H and T:

- Hs must be as explicit as possible to reduce ambiguities and facilitate their correct interpretation;
- Hs must remain as concise as possible, to maintain linguistic “fluency”;
- Hs 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 Hs 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.

4.2.2 The Final Data Set

The Development Set is composed of 10 topics and contains globally 80 Hs and 2,538 sentences. Each sentence of a topic was annotated against each H of the same topic, yielding 20,104 sentence annotations, of which 810 are “entailment” judgments.

As regards the Test Set, originally it should have contained the same number of topics as the

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

Development Set. However, due to inter-annotator agreement problems, one topic had to be removed. The final Test Set is thus composed of 9 topics and contains 81 Hs and 1,949 sentences. The number of sentence annotations amounts to 17,280 and the "entailing" judgments are 800.

In order to assure the creation of a high quality resource, the whole Search 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[9]. After the reconciliation phase, the inter-annotator agreement calculated using the Kappa statistics (Siegel and Castellan, 1988; Fleiss, 1971) was 97.10% for the Development Set and 97.02% for the Test Set[10].

4.3 Evaluation Measures

System results have been 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 have been made available to participants, for both topics and hypotheses. As systems were not forced to retrieve at least one entailing sentence for each topic/hypothesis, in order to calculate macro-averaged results it was decided that, if no sentence was returned for a given topic/hypothesis, the Precision for that topic/hypothesis is 0.

4.4 Submitted Systems and Results

Eight teams participated in the Search Task, submitting a total of 20 runs. Table 10 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 TAC Proceedings.

Some general tendencies can be gathered from the results. First, macro-averaged scores are higher than micro-averaged ones, with the best performances obtained for macro-average over hypotheses. Then, Recall values are on average higher than Precision values. Moreover, the difference between Precision and Recall within each run shows a great variability between the systems, ranging (on micro-averaged results) from 4.94 for *unimelb1* to 75.35 for *Sagan1*.

As regards overall results on micro-average, Table 9 shows some F-measure statistics, calculated both over all the submitted runs and considering only the best run of each participating group.

F-measure	**All runs**	**Best runs**
Highest	45.59	45.59
Lowest	9.55	17.51
Median	30.14	30.2
Average	29.17	30.51

Table 9. Search Task F-measure statistics

A pure Information Retrieval baseline was also calculated, using a search engine to retrieve from each topic corpus the entailing sentences for each hypothesis. The baseline was created taking into account each topic separately, and considering (i) each topic hypothesis as a query, and (ii) the corpus sentences as "the documents" to be retrieved for each query. To this purpose, the Apache Lucene[11] text search engine, Version 2.9.1, was used with the following characteristics:

- *StandardAnalyzer* (tokenization, lower-case and stop-word filtering, basic clean-up of words)
- Boolean "OR" query
- Default document scoring function

Four baselines were calculated, considering as entailing sentences for each H respectively the first 5/10/15/20 sentences top-ranked by Lucene for that given H. The results in Table 11 show

[9] For a detailed analysis of disagreement, see (Bentivogli et al., 2009).

[10] It is worth mentioning that the percentage of agreement over those annotations where at least one assessor said YES was 92% for the Development Set and 91.83% for the Test Set.

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

RUN	Micro-Average			Macro-Average					
				By TOPIC			By HYPOTHESIS		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
BIU1	37.03	55.50	44.42	41.86	53.26	46.88	48.51	59.80	53.57
BIU2	40.49	47.88	43.87	43.72	44.74	44.22	48.96	52.26	50.56
BIU3	40.98	51.38	**45.59**	43.76	48.48	46.00	49.60	55.98	52.60
Boeing1	61.54	15.00	24.12	51.04	13.04	20.77	31.27	13.22	18.58
Boeing2	33.39	25.12	**28.67**	34.96	25.06	29.19	41.19	28.87	33.95
Boeing3	36.41	09.38	14.91	47.24	09.02	15.15	25.77	09.64	14.04
clr091	20.74	42.50	27.88	20.94	41.92	27.93	31.29	46.68	37.47
clr092	20.34	49.25	**28.79**	20.92	48.09	29.16	32.64	52.97	40.39
FBKirst1	24.55	46.50	32.14	26.59	44.33	33.24	35.53	49.79	41.47
FBKirst2	22.54	64.75	**33.44**	27.44	63.50	38.32	33.34	67.95	44.73
FBKirst3	21.78	64.75	32.60	27.48	63.71	38.40	33.02	67.81	44.41
Sagan1	10.16	85.50	**18.16**	10.90	88.06	19.40	12.12	87.87	21.30
ssl1	11.49	36.75	**17.51**	13.28	35.62	19.35	17.52	39.15	24.21
ssl2	10.47	27.25	15.12	11.93	27.59	16.66	12.81	31.78	18.26
ssl3	19.03	06.38	09.55	35.45	07.84	12.84	18.02	09.45	12.39
UAIC20091	51.12	22.75	31.49	53.15	24.00	33.07	46.65	26.01	33.40
UAIC20092	51.12	22.88	**31.61**	53.03	24.08	33.12	46.55	26.42	33.71
unimelb1	42.94	38.00	**40.32**	41.98	30.76	35.50	31.40	37.95	34.37
unimelb2	29.30	44.50	35.33	27.36	37.27	31.56	21.61	45.72	29.35
unimelb3	19.48	48.62	27.82	18.89	42.08	26.08	18.10	49.13	26.45

Table 10. The Search task results (in bold Best run of each system)

that baseline_10, scoring an F-measure of 47.2, performed best, not only with respect to the other baselines, but also with respect to the RTE participating systems. These results suggest that further research about Textual Entailment performed against a real corpus is actually needed; however, given the novelty and the difficulties inherent to the Search Task, we consider both the number of participants and the obtained results as satisfactory.

	Precision	Recall	F-measure
Baseline_5	60.00	30.38	40.33
Baseline_10	46.91	47.5	47.2
Baseline_15	37.98	57.25	45.66
Baseline_20	32.23	64.25	42.92

Table 11. IR baselines results

5 Conclusions and Future Work

The RTE-5 Challenge has demonstrated once again that textual entailment recognition represents an important field of investigation in NLP.

This year's innovations have been successfully introduced in the campaign. As far as the Main Task is concerned, longer and un-edited texts were introduced in the data set in order to make the task more challenging. The stable performances obtained demonstrate that the systems are robust enough to cope with the increasing difficulty of the data sets.

The introduction of the Search Pilot Task put into practice a real interaction between the RTE task and the Summarization task, allowing the analysis of the potential impact of textual entailment recognition on a real NLP application task. The number of participants in the Pilot Task showed that the textual entailment field is mature enough to move towards a more realistic scenario. With respect to the traditional setting, the Search Task imposes new challenges, as systems have to deal with the natural distribution of entailing vs. non-entailing texts and have to face the problem of inference from a complete discourse, even across documents.

Finally, the introduction of ablation tests aimed at evaluating the knowledge resources used by RTE systems had a very positive response. The number of submitted ablation tests and the variety of the resources evaluated confirmed the widespread interest of the community in such resources. This successful experiment represents a first step towards the definition of a new pilot task focused on knowledge resource evaluation, which will be proposed in the next campaign.

If much has been achieved in these five years, there is still more to do in the future. Some interesting approaches have been proposed so far, but still there seems to be room for improvement, as the average performances of the systems show. A more detailed analysis of the types of entailment proposed in the different competitions could provide some useful suggestions on how to improve the data collection and the preparation of the final test set. Finally, the introduction of some metrics which more specifically evaluate the system performances in the three-way task, or which give greater importance to more difficult pairs, could contribute to a more comprehensive analysis of the results.

Acknowledgments

We are grateful to Lance Ramshaw, from BBN, for running his IE system on Wikinews data and making the output available for the production of pairs in the IE setting.

We would also like to acknowledge the people and organizations involved in creating and annotating the data: Leda Casanova, Anna Fioravanti, Pamela Forner, and Medea Lo Leggio from CELCT; Karolina Owczarzak and multiple assessors 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. 2009. Considering Discourse References in Textual Entailment Annotation, In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon (GL 2009)*, Pisa, Italy.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognizing Textual Entailment Challenge. In Quiñonero-Candela et al., editors, *MLCW 2005*, LNAI Volume 3944, pages 177-190. Springer-Verlag.

Hoa T. Dang, and K. Owczarzak. 2008. Overview of the TAC 2008 Update Summarization Task. In *TAC 2008 Proceedings*. http://www.nist.gov/tac/publications/2008/papers.html

Hoa T. Dang, and K. Owczarzak. 2009. Overview of the TAC 2009 Update Summarization Task. In *TAC 2009 Proceedings*. http://www.nist.gov/tac/publications/2009/papers.html

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).

Yashar Mehdad and Bernardo Magnini. 2009. A Word Overlap Baseline for the Recognizing Textual Entailment Task, http://hlt.fbk.eu/sites/hlt.fbk.eu/files/baseline.pdf.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. Evaluating Content Selection in Summarization: The Pyramid Method, in *ACM*

Transactions on Speech and Language Processing, Volume 4 ,Issue 2 (May 2007).

Sidney Siegel and N. John Castellan. 1988. *Non-parametric Statistics for the Behavioural Sciences,* McGraw-Hill, New York.