The Fourth PASCAL Recognizing Textual Entailment Challenge

Danilo Giampiccolo
CELCT
Via alla Cascata 56/c
38100 POVO TN
giampiccolo@celct.it

Hoa Trang Dang
NIST
Gaithersburg, MD, USA
hoa.dang@nist.gov

Bernardo Magnini
FBK-ITC
Via Sommarive 18,
38100 Povo TN
magnini@itc.it

Ido Dagan
Computer Science Department
Bar-Ilan University
Ramat Gan 52900, Israel
dagan@macs.biu.ac.il

Bill Dolan
Microsoft Research
Redmond, WA, 98052, USA
billdol@microsoft.com

Abstract

In 2008 the Recognizing Textual Entailment Challenge (RTE-4) was proposed for the first time as a track at the Text Analysis Conference (TAC). Another important innovation introduced in this campaign was a three-judgment task, which required the systems to make a further distinction between pairs where the entailment does not hold because the content of H is contradicted by the content of T, and pairs where the entailment cannot be determined because the truth of H cannot be verified on the basis of the content of T. A classic two-way task was also proposed. 45 runs were submitted by 26 participants, half of whom chose the 3-way task. This paper describes the preparation of the dataset, and gives an overview of the results achieved by the participating systems.

1 Recognizing Textual Entailment Challenges

1.1 Introduction

Since 2004, the RTE Challenges have promoted research about Textual Entailment Recognition as a generic task that captures major semantic inference needs across many natural language processing applications, such as Question Answering (QA), Information Retrieval (IR), Information Extraction (IE), and multi-document summarization (SUM), providing a common solution for modeling language variability. The RTE task consists in 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. The system is tested against an annotated dataset which includes typical examples that correspond to success and failure cases of the above-mentioned applications. The examples represent different levels of entailment reasoning, such as lexical, syntactic, morphological and logi-

cal. The Textual Entailment Recognition task has raised increasing interest in the NLP community, as it seems to work as a common framework in which to analyse, compare and evaluate different techniques used in NLP applications to deal with semantic inference, a common issue shared by many NLP applications. In fact, by focusing on textual entailment recognition, the RTE task permits a better understanding of the problems related to finding equivalences and similarities at lexical, syntactic and semantic levels. The fact that RTE has recently been the subject of several studies and has been greatly discussed in several papers and workshops all over the world, confirms the impression that there is still a call for further investigation in this field.

1.2 The Previous RTE Challenges

The RTE task was proposed for the first time in 2005 as a challenge for systems which dealt with the typical issues concerning the recognition of textual entailment in different scenarios. This was an area of NLP research which had called for thorough investigation for some time, and it immediately received great attention at a dedicated workshop in Southampton, in April 2005. Building on the success of the first round, a second challenge was organized the following year, and the results were presented at the PASCAL Challenges Workshop in April 2006 in Venice. The event confirmed the successful trend of the previous challenge, recording a growth both in the number of participants and in the quality of the contribution to the research in textual entailment.

As the interest in the task had been constantly increasing during the first two campaigns, it was decided to expand the scope of the RTE challenge, proposing it to a wider audience at a bigger conference. The results of the RTE-3 Challenge were therefore presented in a workshop dedicated to Textual Entailment and Paraphrasing at ACL 2007 in Prague. Although the task maintained the basic structure of the previous campaigns, in order to facilitate the participation of newcomers, some innovations were introduced, such as longer texts so as to stimulate the need for discourse analysis. A pilot task, called "Extending the Evaluation of Inference Texts" was also proposed, which required the participating systems 1) to give a more detailed judgment, making a three-way decision between "YES", "NO", and "UKNOWN" against the same test set used in the main task; and 2) to provide justifications for decisions taken.

The response in terms of participants was very positive this time too, as 26 groups submitted 44 runs, using different approaches and achieving higher scores than in previous challenges. The interest in RTE has been constantly confirmed by the vast scientific production on this subject, and by the fact that RTE modules have been used in real NLP applications, such as the Neumann and Wang's Answer Validation system presented at QA@CLEF 2008 [Neumann 2008].

1.3 The Fourth Challenge

Capitalizing on the positive feedback which RTE has received from the NLP community so far, it was decided that the fourth challenge should make a step forward by introducing some new elements, in order to keep it both affordable and stimulating for the largest number possible of NLP researchers.

A first innovation was represented by the decision to join the efforts of the National Institute of Standard and Technology, which had proposed the pilot task in the 2007 campaign, and CELCT, which had taken part in the organization of all the previous campaigns, presenting RTE-4 as a track of the Text Analysis Conference (http://www.nist.gov/tac/). The goal of this new consortium was to put together the resources of the two institutions in order to reach a larger number of people and, at the same time, to offer better quality in the preparation of the task and in the analysis of the results.

The other major innovation introduced this year was the introduction of the 3 way judgment (already proposed as a pilot task in 2007) in the main task.

For the rest, the basic structure of the challenge remained unchanged. The settings from which the pairs were extracted were Information Extraction, Information Retrieval, Question Answering and Summarization, like in the previous two campaigns, and the same sources and procedures were used for the production of the pairs, in an attempt to facilitate the comparison between the performances of systems which had participated in the previous campaigns.

TASK	TEXT	HYPOTHESIS	ENTAILMENT
IE	Admiral Kuroyedov was in charge of the navy during the Kursk disaster of 2000, in which 118 sailors died when their submarine sank. Kuroyedov is being replaced by Vladimir Masorin, who was previously serving as the Chief of staff for the Russian Navy.	Kuroyedov caused the Kursk disaster.	UNKNOWN
IE	Spencer Dryden, the drummer of the legendary American rock band Jefferson Airplane, passed away on Tuesday, Jan. 11. He was 66. Dryden suffered from stomach cancer and heart disease.	Spencer Dryden died at 66.	ENTAILMENT
IR	The Dalai Lama today called for Tibetans to end protests against the Beijing Olympics, also telling MPs in London he would happily accept an invitation to attend the event if relations with China improved.	China hosts Olympic games.	ENTAILMENT
IR	Lower food prices pushed the UK's inflation rate down to 1.1% in August, the lowest level since 1963. The headline rate of inflation fell to 1.1% in August, pushed down by falling food prices.	Food prices are on the increase.	CONTRADICTION
QA	The gambusia affinis, dubbed the mosquito fish, is an aquatic predator that devours mosquito larvae. Officials are releasing the fish into the fetid waters of abandoned pools to reduce the burgeoning mosquito population.	Gambusia is a species of mosquito.	CONTRADICTION
QA	Four people were killed and at least 20 injured when a tornado tore through an Iowa boy scout camp on Wednesday, where dozens of scouts were gathered for a summer retreat, state officials said.	Four boy scouts were killed by a tornado.	UNKNOWN
SUM	Kingdom flag carrier British Airways (BA) has entered into merger talks with Spanish airline Iberia Lineas Aereas de Espana SA. BA is already Europe's third-largest airline.	The Spanish airline Iberia Lineas Aereas de Espana SA is Europe's third-largest airline.	CONTRADICTION

Table 1. Examples taken from the test set.

2 The RTE-4 Dataset

2.1 Task description and dataset overview

The participating systems were assigned the task of recognizing textual entailment in a set of 1000 T-H pairs; i.e., they were required to decide, given a set of text pairs, called *T(ext)* and *H(ypothesis)*, whether T entailed H or not. Textual entailment is defined as a directional relation between two text fragment – 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*.
Unlike in the previous campaigns, the main task asked the systems to make a three-way decision, further distinguishing, in case there was no entailment between T and H, whether the truth of H was contradicted by T, or remained unknown on the basis of the information contained in T.
In other words, the participating systems had to decide whether:

- T entailed H - in which case the pair was marked as ENTAILMENT
- T contradicted H - in which case the pair was marked as CONTRADICTION
- The truth of H could not be determined on the basis of T - in which case the pair was marked as UNKNOWN

The classic two-way RTE task was also offered, in which the pairs where T entailed H were marked as ENTAILMENT, and those where the entailment did not hold were marked as NO ENTAILMENT.
No development set was provided this year, as the pairs proposed were very similar to the ones contained in last year's development and test sets, which could therefore be used to train the systems.
Four applications – namely IE, IR, QA and SUM – were considered as settings or contexts for the pairs generation (see below for a detailed description). The length of the H's was the same as in the past datasets; however, the T's were generally longer, following the decision taken last year of moving towards real cases where more discourse analysis is required. A major difference with respect to previous campaigns was that the RTE-4 dataset consisted of 1000 T-H pairs, instead of 800. This was due to the fact that while 200 pairs were selected for QA and SUM, 300 were chosen for IE and IR, as these two settings proved somewhat more difficult in the previous campaigns.
The distribution according to the 3 way annotation, both in the individual settings and in the overall test set, was as follows:

- 50% ENTAILMENT
- 35% UNKNOWN
- 15% CONTRADICTION.

2.2 Pair Collection

As usual, human annotators generated T-H pairs within the four aforementioned application settings, following exactly the same process as used last year.

The IE task was inspired by the Information Extraction and Relation Extraction applications, simulating the need of an IE system to recognize that the given text indeed entails the semantic relation that is expected to hold between the candidate template slot fillers.

The T-H pairs which replaced text and structure templates of the IE task were produced in the following ways:

1. 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.
2. The same material was used and the news articles were also used to manually generate entailment pairs based on ACE relations, simulating the extraction process performed by IE systems.
3. New relations, such as "X discover Y", "X win Y", etc., were considered both to be processed by IE systems and to manually generate T-H pairs from collected news articles.

In the IR (Information Retrieval) application setting, the hypotheses were propositional IR queries, e.g. "corn prices increase". Texts that did or did

not entail the hypotheses were selected from documents retrieved by different search engines such as Google, Yahoo and MSN, for each hypothesis. In this application setting, the given propositional hypotheses are assumed to be entailed by relevant retrieved documents.

For the QA (Question Answering) setting, both questions taken from the datasets of official QA competitions, such as TREC QA and QA@CLEF datasets, and questions produced specifically for the purposes of RTE were fed to actual QA systems, which retrieved answers from the Web. Then, human annotators transformed the question-answer pairs into T-H pairs as follows:

- An answer term of the expected answer type was picked from the answer passage - either a correct or an incorrect one.
- The question was turned into an affirmative sentence plugging in the answer term.
- H-T pairs were generated, using the affirmative sentences as hypotheses (H's) and the original answer passages as texts (T's).
- For example, given the question "*How many seconds did it take Tyson Gay to run 100 meters?*" and a text "*When Tyson Gay crossed the finish line in the men's 100 meters yesterday, the crowd at Hayward Field gasped. The clock displayed 9.68 seconds. Everyone at the US Olympic track and field trials knew what that meant.*", the piece of information "*9.68 seconds*" was extracted from the text and inserted into the question, which was finally turned into the declarative sentence "*Tyson Gay ran 100 meters in 9.68 seconds*", which became the hypothesis of a pair where the entailment held.
- Examples for which the entailment did not hold were created by producing H's where the piece of information answering the implied question was not relevant or contradicted the content of the T.

Using the RTE process, QA systems can verify that the retrieved passage text actually entails the provided answer (see Ave Exercise, Rodrigo 2008).

In the SUM (Summarization) setting, T'*s* and H'*s* were sentences taken from a news document cluster, a collection of news articles that describe the same news item. Annotators were given the output of multi-document summarization systems - including the document clusters and the summary generated for each cluster. Then they picked sentence pairs with high lexical overlap, preferably where at least one of the sentences was taken from the summary (this sentence usually played the role of T). For positive examples, the hypothesis was simplified by removing sentence parts, until it was fully entailed by T. Negative examples, where the entailment did not hold, were produced in a similar way, i.e. taking away parts of T so that the final information contained in H either contradicted the content of T, or was not enough to determine whether T entailed H.

2.3 The final dataset

As in previous challenges, each pair of the dataset was judged by three annotators. Pairs on which the annotators disagreed were discarded. The disagreement between annotators was often due to the fact that one annotator did not consider that some extra information was contained in the H, making it more specific than the T. In other cases, the disagreement was about whether the information in H was contradictory with respect to the content of T, or simply not sufficient to determine a judgment, especially in some ambiguous cases. A number of pairs were also discarded because they were too similar to others, or their content was otherwise inappropriate.

Both texts and hypotheses were revised by native English speakers to eliminate the major spelling and grammar mistakes frequently present in texts taken from the web. No major changes were otherwise made, in order to keep the exercise realistic.

2.4 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.

First Author	3-way Task			2-way Task		First Author	3-way Task			2-way Task	
	Accuracy 3-way judgment	Accuracy 2-way judgment	Average precision	Accuracy	Average precision		Accuracy 3-way judgment	Accuracy 2-way judgment	Average precision	Accuracy	Average precision
AUEBNLP1	0.546	0.58	0.5654	0.566	0.5464	IIITSum081	0.309	0.531			
AUEBNL2	0.547	0.579	0.562	0.578	0.563	IIITSum082	0.307	0.529			
AUEBNL3	0.554	0.584	0.522	0.566	0.5465	IPD1	0.427	0.512			
BIU1				0.583		IPD2	0.414	0.51			
BIU2				0.573		IPD3	0.432	0.54			
BIU3				0.584		LCC				0.746	0.7419
boeing1	0.377	0.515				PeMoZa1				0.563	0.5619
boeing2	0.459	0.565				PeMoZa2				0.59	0.6287
boeing3	0.481	0.547				PeMoZa3				0.586	0.603
cambridge1				0.51	0.5242	QUANTA1	0.588	0.633	0.6332	0.659	0.6225
cambridge2				0.516	0.5257	QUANTA2				0.623	0.5926
CERES1	0.405	0.526		0.521		OAQA1	0.616	0.688	0.5811		
CERES2	0.416	0.526				OAQA2	0.52	0.54	0.5811		
CLEAR1				0.595	0.6092	OAQA3	0.432	0.547	0.581		
CLEAR2				0.603	0.613	Sagan1	0.538	0.576			
CLEAR3				0.606	0.6254	Sagan2	0.546	0.571			
DFKI1	0.614	0.687		0.672		Stanford1	0.553	0.614	0.4416		
DFKI2	0.606	0.67		0.699		Stanford2	0.508	0.57	0.5427		
DFKI3	0.56	0.664		0.706		Stanford3	0.508	0.57	0.5427		
DLSIUAES1				0.608		UAIC20081	0.685	0.72		0.721	
DLSIUAES2				0.599		UNED1				0.549	
DLSIUAES3				0.594		UNED2				0.513	
Emory1	0.546	0.576	0.5994	0.588	0.5998	UNED3				0.54	
Emory2	0.547	0.583	0.5954	0.57	0.6012	uoeltg1				0.582	
Emory3	0.543	0.564	0.5975	0.511	0.5115	Uoeltg2				0.57	
fbkirst1				0.54	0.4946	Uoeltg3				0.524	
fbkirst2				0.546	0.5516	UPC1				0.563	
fbkirst3				0.57	0.553	UPC2				0.554	
FSC1	0.466	0.526				UPC3				0.561	
FSC2				0.526		UMD1	0.556	0.619	0.4427		
KUNLP1				0.51		UMD2	0.556	0.617	0.4426		
KUNLP2				0.519		UMD3	0.554	0.617	0.4408		
KUNLP3				0.497		wlvUK1				0.571	

Table 2. Submission results

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 re-tagged as NO ENTAILMENT.

3 Submitted systems and results

Twenty-six teams participated in the fourth RTE challenge, the same number as last year: 13 from Europe, 9 from America, and 4 from Asia. They submitted 45 runs overall. Most of them belonged to the academic world (21), but research centers (4) and industrial companies (1) were also represented.

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, 8 participants took part only in the 3-way task; 13 only in the 2-way task, and 5 in both - which means that half of the participants chose to test their systems against the three-way judgment task.

As regards the results, in the 3-way task, the best accuracy was 0.685 calculated against the 3-way judgment, and 0.72 calculated against the 2-way judgment. The 3-way task appeared to be altogether quite challenging, as the average 3-way score was 0.51, quite low compared to the results achieved in previous campaigns. The systems performed better in the 2-way task, achieving accuracy scores which ranged between 0.459 and 0.746, with an average score of 0.573. These results are lower than those achieved in last year's competition, where the accuracy scores ranged from 0.49 to 0.80, even though a comparison is not really possible as the datasets were actually different.

As a general remark, the IE setting appeared to be the more difficult, recording the lowest accuracy scores; meanwhile SUM and IR settings seemed to be easier. As this trend was present also in the previous challenges, focusing on the analysis of the differences between settings could help improve the understanding of inference phenomena and therefore stimulate advances also in the research about the RTE task itself.

4 Conclusions and future work

The Fourth RTE Challenge has demonstrated once again that textual entailment recognition represents an important field of investigation in NLP. In fact, it attracted a considerable number of participants, testifying that many researchers are still interested in studying more in depth a task that

appears to address a core issue, common to many different applications.

The innovations introduced this year, which were aimed at keeping the competition both feasible for new-comers and stimulating for "veterans", seem to have reached their goal. The collaboration between NIST and CELCT made the organization easier and contributed to enlarge the scope of the challenge, involving new subjects in the competition and proposing the task in the new scenario of the Text Analysis Conference. On the other hand, the introduction of the three-way judgments in the main task made the exercise both more realistic and challenging, as the systems had to make a further distinction between pairs where the entailment did not hold due to contradiction and pairs where it was not possible to give either a positive or negative judgment due to lack of information.

If much has been achieved in these four years, there is still more to do in the future. Some interesting approaches have been proposed so far, but there seems to be still room for improvement, as the average performances of the systems showed. A deeper analysis of the results, considering also the approaches used by the different systems, could help to understand what methodologies have proved to offer the best solutions. At the same time, a comparative study of the datasets so far produced, and 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 evaluated performance in the three-way task, or which gave greater importance to more difficult test pairs, could contribute to a more comprehensive analysis of the results.

Acknowledgments

The following sources were used in the preparation of the data:

• PowerAnswer question answering system, from Language Computer Corporation, provided by Dan Moldovan and Marta Tatu.
http://www.languagecomputer.com/solutions/question answering/power answer/

• Cicero Custom and Cicero Relation information extraction systems, from Language Computer Corporation, provided by Sanda M. Harabagiu, Andrew Hickl, John Lehmann and Paul Aarseth.
http://www.languagecomputer.com/solutions/information_extaction/cicero/index.html

• Columbia NewsBlaster multi-document summarization system, from the Natural Language Processing group at Columbia University's Department of Computer Science.
http://newsblaster.cs.columbia.edu/

• NewsInEssence multi-document summarization system provided by Dragomir R. Radev and Jahna Otterbacher from the Computational Linguistics and Information Retrieval research group, University of Michigan.
http://www.newsinessence.com

• New York University's information extraction system, provided by Ralph Grishman, Department of Computer Science, Courant Institute of Mathematical Sciences, New York University.

• TREC IR queries and TREC-QA question collections, from the National Institute of Standards and Technology (NIST).
http://trec.nist.gov/

• CLEF IR queries and CLEF-QA question collections, from DELOS Network of Excellence for Digital Libraries.
http://www.clef-campaign.org/, http://clef-qa.itc.it/

We would like to thank the people and organizations that made these sources available for the challenge. In addition, we thank Idan Szpektor and Roy Bar Haim from Bar-Ilan University for their assistance and advice.

We would also like to acknowledge the people and organizations involved in creating and annotating the data: Pamela Forner, Errol Hayman, Leda Casanova and Valentina Bruseghini from CELCT;
Karolina Owczarzak, Ellen Voorhees, and multiple assessors from NIST.

This work was 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.

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.

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.

Rui Wang and Günter Neumann. 2008. Overview of the Answer Validation Exercise 2008, CLEF 2008 Working Notes, http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html.

Álvaro Rodrigo, Anselmo Peñas, Felisa Verdejo. 2008. Overview of the Answer Validation Exercise 2008, CLEF 2008 Working Notes, http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html.