The Third PASCAL Recognizing Textual Entailment Challenge

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

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

This paper presents the Third PASCAL Recognising Textual Entailment Challenge (RTE-3), providing an overview of the dataset creating methodology and the submitted systems. In creating this year's dataset, a number of longer texts were introduced to make the challenge more oriented to realistic scenarios. Additionally, a pool of resources was offered so that the participants could share common tools. A pilot task was also set up, aimed at differentiating unknown entailments from identified contradictions and providing justifications for overall system decisions. 26 participants submitted 44 runs, using different approaches and generally presenting new entailment models and achieving higher scores than in the previous challenges.

1.1 The RTE challenges

The goal of the RTE challenges has been to create a benchmark task dedicated to textual entailment – recognizing that the meaning of one text is entailed, i.e. can be inferred, by another[1]. In the recent years, this task has raised great interest since applied semantic inference concerns many practical Natural Language Processing (NLP) applications, such as Question Answering (QA), Information Extraction (IE), Summarization, Machine Translation and Paraphrasing, and certain types of queries in Information Retrieval (IR). More specifically, the RTE challenges have aimed to focus research and evaluation on this common underlying semantic inference task and separate it from other problems that different NLP applications need to handle. For example, in addition to textual entailment, QA systems need to handle issues such as answer retrieval and question type recognition.

By separating out the general problem of textual entailment from these task-specific problems, progress on semantic inference for many application areas can be promoted. Hopefully, research on textual entailment will finally lead to the development of entailment "engines", which can be used as a standard module in many applications (similar to the role of part-of-speech taggers and syntactic parsers in current NLP applications).

In the following sections, a detailed description of RTE-3 is presented. After a quick review

[1] The task was first defined by Dagan and Glickman (2004).

Proceedings of the Workshop on Textual Entailment and Paraphrasing, pages 1–9,
Prague, June 2007.

of the previous challenges (1.2), section 2 describes the preparation of the dataset. In section 3 the evaluation process and the results are presented, together with an analysis of the performance of the participating systems.

1.2 The First and Second RTE Challenges

The first RTE challenge[2] aimed to provide the NLP community with a new benchmark to test progress in recognizing textual entailment, and to compare the achievements of different groups. This goal proved to be of great interest, and the community's response encouraged the gradual expansion of the scope of the original task.

The Second RTE challenge[3] built on the success of the first, with 23 groups from around the world (as compared to 17 for the first challenge) submitting the results of their systems. Representatives of participating groups presented their work at the PASCAL Challenges Workshop in April 2006 in Venice, Italy. The event was successful and the number of participants and their contributions to the discussion demonstrated that Textual Entailment is a quickly growing field of NLP research. In addition, the workshops spawned an impressive number of publications in major conferences, with more work in progress. Another encouraging sign of the growing interest in the RTE challenge was represented by the increase in the number of downloads of the challenge datasets, with about 150 registered downloads for the RTE-2 development set.

1.3 The Third Challenge

RTE-3 followed the same basic structure of the previous campaigns, in order to facilitate the participation of newcomers and to allow "veterans" to assess the improvements of their systems in a comparable test exercise. Nevertheless, some innovations were introduced, on the one hand to make the challenge more stimulating and, on the other, to encourage collaboration between system developers. In particular, a limited number of longer texts, i.e. up to a paragraph in length, were incorporated in order to move toward more comprehensive scenarios, which incorporate the need for discourse analysis. However, the majority of examples remained similar to those in the previous challenges, providing pairs with relatively short texts.

Another innovation was represented by a resource pool[4], where contributors had the possibility to share the resources they used. In fact, one of the key conclusions at the second RTE Challenge Workshop was that entailment modeling requires vast knowledge resources that correspond to different types of entailment reasoning. Moreover, entailment systems also utilize general NLP tools such as POS taggers, parsers and named-entity recognizers, sometimes posing specialized requirements to such tools. In response to these demands, the RTE Resource Pool was built, which may serve as a portal and forum for publicizing and tracking resources, and reporting on their use.

In addition, an optional pilot task, called "*Extending the Evaluation of Inferences from Texts*" was set up by the US National Institute of Standards and Technology (NIST), in order to explore two other sub-tasks closely related to textual entailment: differentiating unknown entailments from identified contradictions and providing justifications for system decisions. In the first sub-task, the idea was to drive systems to make more precise informational distinctions, taking a three-way decision between "YES", "NO" and "UNKNOWN”, so that a hypothesis being unknown on the basis of a text would be distinguished from a hypothesis being shown false/contradicted by a text. As for the other subtask, the goal for providing justifications for decisions was to explore how eventual users of tools incorporating entailment can be made to understand how decisions were reached by a system, as users are unlikely to trust a system that gives no explanation for its decisions. The pilot task exploited the existing RTE-3 Challenge infrastructure and evaluation process by using the same test set, while utilizing human assessments for the new sub-tasks.

[2] http://www.pascal-network.org/Challenges/RTE/.
[3] http://www.pascal-network.org/Challenges/RTE2./

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

TASK	TEXT	HYPOTHESIS	ENTAILMENT
IE	At the same time the Italian digital rights group, Electronic Frontiers Italy, has asked the nation's government to investigate Sony over its use of anti-piracy software.	Italy's government investigates Sony.	NO
IE	Parviz Davudi was representing Iran at a meeting of the Shanghai Co-operation Organisation (SCO), the fledgling association that binds Russia, China and four former Soviet republics of central Asia together to fight terrorism	China is a member of SCO.	YES
IR	Between March and June, scientific observers say, up to 300,000 seals are killed. In Canada, seal-hunting means jobs, but opponents say it is vicious and endangers the species, also threatened by global warming	Hunting endangers seal species.	YES
IR	The Italian parliament may approve a draft law allowing descendants of the exiled royal family to return home. The family was banished after the Second World War because of the King's collusion with the fascist regime, but moves were introduced this year to allow their return.	Italian royal family returns home.	NO
QA	Aeschylus is often called the father of Greek tragedy; he wrote the earliest complete plays which survive from ancient Greece. He is known to have written more than 90 plays, though only seven survive. The most famous of these are the trilogy known as Orestia. Also well-known are The Persians and Prometheus Bound.	"The Persians" was written by Aeschylus.	YES
SUM	A Pentagon committee and the congressionally chartered Iraq Study Group have been preparing reports for Bush, and Iran has asked the presidents of Iraq and Syria to meet in Tehran.	Bush will meet the presidents of Iraq and Syria in Tehran.	NO

Table 1: Some examples taken from the Development Set.

2 The RTE-3 Dataset

2.1 Overview

The textual entailment recognition task required the participating systems to decide, given two text snippets *t* and *h*, whether *t* entails *h*. Textual entailment is defined as a directional relation between two text fragments, called *text* (*t*, the entailing text), and hypothesis (*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*.

As in the previous challenges, the RTE-3 dataset consisted of 1600 text-hypothesis pairs, equally divided into a development set and a test set. While the length of the hypotheses (*h*) was the same as in the past datasets, a certain number of texts (*t*) were longer than in previous datasets, up to a paragraph. The longer texts were marked as L, after being selected automatically when exceeding 270 bytes. In the test set they were about 17% of the total.

As in RTE-2, four applications – namely IE, IR, QA and SUM – were considered as settings or contexts for the pairs generation (see 2.2 for a detailed description). 200 pairs were selected for each application in each dataset. Although the datasets were supposed to be perfectly balanced, the number of negative examples were slightly higher in both development and test sets (51.50% and 51.25% respectively; this was unintentional). Positive entailment examples, where *t* entailed *h*, were annotated *YES*; the negative ones, where entailment did not hold, *NO*. Each pair was annotated with its related task (IE/IR/QA/SUM) and entailment judgment (YES/NO, obviously released only in the development set). Table 1 shows some examples taken from the development set.

The examples in the dataset were based mostly on outputs (both correct and incorrect) of Web-based systems. In order to avoid copyright problems, input data was limited to either what had already been publicly released by official competitions or else was drawn from freely available sources such as WikiNews and Wikipedia.

In choosing the pairs, the following judgment criteria and guidelines were considered:

- As entailment is a directional relation, the hypothesis must be entailed by the given text, but the text need not be entailed by the hypothesis.
- The hypothesis must be fully entailed by the text. Judgment must be NO if the hypothesis includes parts that cannot be inferred from the text.
- Cases in which inference is very probable (but not completely certain) were judged as YES.
- Common world knowledge was assumed, e.g. the capital of a country is situated in that country, the prime minister of a state is also a citizen of that state, and so on.

2.2 Pair Collection

As in RTE-2, human annotators generated *t-h* pairs within 4 application settings.

The IE task was inspired by the Information Extraction (and Relation Extraction) application, where texts and structured templates were replaced by *t-h* pairs. As in the 2006 campaign, the pairs were generated using four different approaches:

1) Hypotheses were taken from the relations tested in the ACE-2004 RDR task, while texts were extracted from the outputs of actual IE systems, which were provided with relevant news articles. Correctly extracted instances were used to generate positive examples and incorrect instances to generate negative examples.
2) The same procedure was followed using output of IE systems on the dataset of the MUC-4 TST3 task, in which the events are acts of terrorism.
3) The annotated MUC-4 dataset and the news articles were also used to manually generate entailment pairs based on ACE relations.
4) Hypotheses corresponding to relations not found in the ACE and MUC datasets were used both to be given to IE systems and to manually generate *t-h* pairs from collected news articles. Examples of these relations, taken from various semantic fields, were "X beat Y", "X invented Y", "X steal Y" etc.

The common aim of all these processes was to simulate the need of IE systems to recognize that the given text indeed entails the semantic relation that is expected to hold between the candidate template slot fillers.

In the IR (Information Retrieval) application setting, the hypotheses were propositional IR queries, which specify some statement, e.g. "*robots are used to find avalanche victims*". The hypotheses were adapted and simplified from standard IR evaluation datasets (TREC and CLEF). Texts (*t*) that did or did not entail the hypotheses were selected from documents retrieved by different search engines (e.g. Google, Yahoo and MSN) for each hypothesis. In this application setting it was assumed that relevant documents (from an IR perspective) should entail the given propositional hypothesis.

For the QA (Question Answering) task, annotators used questions taken from the datasets of official QA competitions, such as TREC QA and QA@CLEF datasets, and the corresponding answers extracted from the Web by actual QA systems. Then they 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.
- *t-h* pairs were generate, using the affirmative sentences as hypotheses (*h's)* and the original answer passages as texts (*t*'s).

For example, given the question "*How high is Mount Everest?*" and a text (*t*) "*The above mentioned expedition team comprising of 10 members was permitted to climb 8848m. high Mt. Everest from Normal Route for the period of 75 days from 15 April, 2007 under the leadership of Mr. Wolf Herbert of Austria*", the annotator, extracting the piece of information "8848m." from the text, would turn the question into an the affirmative sentence "*Mount Everest is 8848m high*", generating a positive entailment pair. This process simulated the need of a QA system to verify that the retrieved passage text actually entailed the provided answer.

In the SUM (Summarization) setting, the entailment pairs were generated using two procedures.

In the first one, *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 were simplified in a similar manner. In alternative, "*pyramids*" produced for the experimental evaluation mehod in DUC 2005 (Passonneau et al. 2005) were exploited. In this new evaluation method, humans select sub-sentential content units (SCUs) in several manually produced summaries on a subject, and collocate them in a "pyramid", which has at the top the SCUs with the higher frequency, i.e. those which are present in most summaries. Each SCU is identified by a label, a sentence in natural language which expresses the content. Afterwards, the annotators individuate the SCUs present in summaries generated automatically (called *peers*), and link them to the ones present in the pyramid, in order to assign each peer a weight. In this way, the SCUs in the automatic summaries linked to the SCUs in the higher tiers of the pyramid are assigned a heavier weight than those at the bottom. For the SUM setting, the RTE-3 annotators selected relevant passages from the peers and used them as T's, meanwhile the labels of the corresponding SCUs were used as H's. Small adjustments were allowed, whenever the texts were not grammatically acceptable. This process simulated the need of a summarization system to identify information redundancy, which should be avoided in the summary.

2.3 Final dataset

Each pair of the dataset was judged by three annotators. As in previous challenges, pairs on which the annotators disagreed were filtered-out.

On the test set, the average agreement between each pair of annotators who shared at least 100 examples was 87.8%, with an average Kappa level of 0.75, regarded as substantial agreement according to Landis and Koch (1997).

19.2 % of the pairs in the dataset were removed from the test set due to disagreement. The disagreement was generally due to the fact that the *h* was more specific than the *t*, for example because it contained more information, or made an absolute assertion where *t* proposed only a personal opinion. In addition, 9.4 % of the remaining pairs were discarded, as they seemed controversial, too difficult, or too similar when compared to other pairs.

As far as the *texts* extracted from the web are concerned, spelling and punctuation errors were sometimes fixed by the annotators, but no major change was allowed, so that the language could be grammatically and stylistically imperfect. The hypotheses were finally double-checked by a native English speaker.

3 The RTE-3 Challenge

3.1 Evaluation measures

The evaluation of all runs submitted in RTE-3 was automatic. The judgments (classifications) returned by the system were compared to the Gold Standard compiled by the human assessors. The main evaluation measure was *accuracy,* i.e. the percentage of matching judgments.

For systems that provided a confidence-ranked list of the pairs, in addition to the YES/NO judgment, an Average Precision measure was also computed. 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). Average precision 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 YES, or, more formally:

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

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

In other words, the more the system was confident that t entails h, the higher was the ranking of the pair. 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.

3.2 Submitted systems

Twenty-six teams participated in the third challenge, three more than in previous year. Table 2 presents the list of the results of each submitted runs and the components used by the systems. Overall, we noticed a move toward deep approaches, with a general consolidation of approaches based on the syntactic structure of Text and Hypothesis. There is an evident increase of systems using some form of logical inferences (at least seven systems). However, these approaches, with few notably exceptions, do not seem to be consolidated enough, as several systems show results not still at the state of art (e.g. Natural Logic introduced by Chambers et al.). For many systems an open issue is the availability and integration of different and complex semantic resources-

A more extensive and fine grained use of specific semantic phenomena is also emerging. As an example, Tatu and Moldovan carry on a sophisticated analysis of named entities, in particular Person names, distinguishing first names from last names. Some form of relation extraction, either through manually built patterns (Chambers et al.) or through the use of an information extraction system (Hickl and Bensley) have been introduced this year, even if still on a small scale (i.e. few relations).

On the other hand, RTE-3 confirmed that both machine learning using lexical-syntactic features and transformation-based approaches on dependency representations are well consolidated techniques to address textual entailment. The extension of transformation-based approaches toward probabilistic settings is an interesting direction investigated by some systems (e.g. Harmeling). On the side of "light" approaches to textual entailment, Malakasiotis and Androutpoulos provide a useful baseline for the task (0.61%) using only POS tagging and then applying string-based measures to estimate the similarity between Text and Hypothesis.

As far as resources are concerned, lexical databases (mostly WordNet and DIRT) are still widely used. Extended WordNet is also a common resource (for instance in Iftene and Balahur-Dobrescu) and the Extended Wordnet Knowledge Base has been successfully used in (Tatu and Moldovan). Verb-oriented resources are also largely present in several systems, including Framenet (e.g. Burchardt et al.), Verbnet (Bobrow et al.) and Propbank (e.g. Adams et al.). It seems that the use of the Web as a resource is more limited when compared to the previous RTE workshop. However, as in RTE-2, the use of large semantic resources is still a crucial factor affecting the performance of systems (see, for instance, the use of a large corpus of entailment examples in Hickl and Bensley).

Finally, an interesting aspect is that, stimulated by the percentage of longer texts included this year, a number of participating systems addressed anaphora resolution (e.g. Delmonte, Bar-Haim et al., Iftene and Balahur-Dobrescu).

3.3 Results

The accuracy achieved by the participating systems ranges from 49% to 80% (considering the best run of each group), while most of the systems obtained a score in between 59% and 66%. One submission, Hickl and Bensley achieved 80% accuracy, scoring 8% higher than the second system (Tatu and Moldovan, 72%), and obtaining the best absolute result achieved in the three RTE challenges.

First Author	Accuracy	Average precision	System Components								
			Lexical Relation, WordNet	n-gram\word similarity	Syntactic Match-ing\Aligning	Semantic Role Labeling\Framenet\Probank, Verbnet	Logical Inference	Corpus/ Web-based Statistics, LSA	ML Classification	Anaphora resolution	Entailment Corpora – DIRT Background Knowledge
Adams	0.6700		X	X				X	X		
Bar-Haim	0.6112	0.6118	X		X			X		X	X
	0.5837	0.6093	X		X			X		X	
Baral	0.4963	0.5364	X				X				X
Blake	0.6050	0.5897	X		X				X		
	0.6587	0.6096	X		X				X		
Bobrow	0.5112	0.5720	X			X	X				
	0.5150	0.5807	X			X	X				
Burchardt	0.6250		X		X	X					
	0.6262										
Burek	0.5500			X				X			
	0.5500	0.5514									
Chambers	0.6050	0.6341	X		X		X		X	X	
	0.6362	0.6527	X		X		X		X	X	
Clark	0.5088	0.4961	X				X				X
	0.4725	0.4961	X				X				X
Delmonte	0.5875	0.5830	X		X	X	X			X	
Ferrandez	0.6563		X	X	X						
	0.6375										
Ferrés	0.6062		X	X					X		
	0.6150		X	X					X		
Harmling	0.5600	0.5813	X		X				X		
	0.5775	0.5952	X		X				X		
Hickl	0.8000	0.8815	X	X			X		X	X	X
Iftene	0.6913		X		X						X
	0.6913		X		X						X
Li	0.6400		X	X					X		
	0.6488										
Litkowski	0.6125										
Malakasiotis	0.6175	0.6808		X					X		
Marsi	0.5913				X						X
Montejo-Ràez	0.5888		X	X	X				X		
	0.6038		X	X	X				X		
Rodrigo	0.6238		X	X	X				X		
	0.6312		X	X	X				X		
Roth	0.6262		X	X							X
	0.5975				X					X	
Settembre	0.6100	0.6195	X	X					X		
	0.6262	0.6274	X	X					X		
Tatu	0.7225	0.6942	X				X			X	X
	0.7175	0.6797	X				X			X	
Wang	0.6650				X				X		
	0.6687										
Zanzotto	0.6675	0.6674	X		X				X		
	0.6575	0.6732	X		X				X		

Table 2: Submission results and components of the systems.

As far as the per-task results are concerned, the trend registered in RTE-2 was confirmed, in that there was a marked difference in the performances obtained in different task settings.

In fact, the average accuracy achieved in the QA setting (0.71) was 20 points higher than that achieved in the IE setting (0.52); the average accuracy in the IR and Sum settings was 0.66 and 0.58 respectively. In RTE-2 the best results were achieved in SUM, while the lower score was always recorded in IE. As already pointed out by Bar-Haim (2006), these differences should be further investigated, as they could lead to a sensible improvement of the performance.

As for the LONG pairs, which represented a new element of this year's challenge, no substantial difference was noted in the systems' performances: the average accuracy over the long pairs was 58.72%, compared to 61.93% over the short ones.

4 Conclusions and future work

At its third round, the Recognizing Textual Entailment task has reached a noticeable level of maturity, as the very high interest in the NLP community and the continuously increasing number of participants in the challenges demonstrate. The relevance of Textual Entailment Recognition to different applications, such as the AVE[5] track at QA at CLEF[6], has also been acknowledged. Furthermore, the debates and the numerous publications about the Textual Entailment have contributed to the better understanding the task and its nature.

To keep a good balance between the consolidated main task and the need for moving forward, longer texts were introduced in the dataset, in order to make the task more challenging, and a pilot task was proposed. The Third RTE Challenge have also confirmed that the methodology for the creation of the datasets, developed in the first two campaigns, is robust. Overall, the transition of the challenge coordination from Bar-Ilan –which organized the first two challenges- to CELCT was successful, though some problems were encountered, especially in the preparation of the data set. The systems which took part in RTE-3 showed that the technology applied to Entailment Recognition has made significant progress, confirmed by the results, which were generally better than last year. In particular, visible progress in defining several new principled scenarios for RTE was represented, such as Hickl's commitment-based approach, Bar Haim's proof system, Harmeling's probabilistic model, and Standford's use of Natural Logic.

If, on the one hand, the success that RTE has had so far is very encouraging, on the other, it incites to overcome certain current limitations, and to set realistic and, at the same time, stimulating goals for the future. First at all, theoretical refinements both of the task and the models applied to it need to be developed. In particular, more efforts are required to improve knowledge acquisition, as little progress has been made on this front so far. Also the data set generation and the evaluation methodology need to be refined and extended. A major problem in the current setting of the data collection is that the distribution of the examples is arbitrary to a large extent, being determined by manual selection. Therefore new evaluation methodologies, which can reflect realistic distributions should be investigated, as well as the possibility of evaluating Textual Entailment Recognition within additional concrete application scenarios, following the spirit of the QA Answer Validation Exercise.

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 and Paul Aarseth.
http://www.languagecomputer.com/solutions/information_ext action/cicero/index.html

• Columbia NewsBlaster multi-document summarization system, from the Natural Language Proc-

[5] http://nlp.uned.es/QA/ave/.
[6] http://clef-qa.itc.it/.

essing group at Columbia University's Departmentof 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.

• MUC-4 information extraction dataset, from the National Institute of Standards and Technology (NIST).
http://www.itl.nist.gov/iaui/894.02/related projects/muc/

• ACE 2004 information extraction templates, from the National Institute of Standards and Technology (NIST).
http://www.nist.gov/speech/tests/ace/

• 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/

• DUC 2005 annotated peers, from Columbia University, NY, provided by Ani Nenkova.
http://www1.cs.columbia.edu/~ani/DUC2005/

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, and Valentina Bruseghini from CELCT for managing the RTE-3 website.

We would also like to acknowledge the people and organizations involved in creating and annotating the data: Pamela Forner, Errol Hayman, Cameron Fordyce from CELCT and Courtenay Hendricks, Adam Savel and Annika Hamalainen from the Butler Hill Group, which was funded by Microsoft Research.

This work was supported in part by the IST Programme of the European Community, under the *PASCAL Network of Excellence*, IST-2002-506778. We wish to thank the managers of the PASCAL challenges program, Michele Sebag and Florence d'Alche-Buc, for their efforts and support, which made this challenge possible. We also thank David Askey, who helped manage the RTE 3 website.

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.

J. R. Landis and G. G. Koch. 1997. The measurements of observer agreement for categorical data. *Biometrics*, 33:159–174.

Rebecca Passonneau, Ani Nenkova., Kathleen McKeown, and Sergey Sigleman. 2005. Applying the pyramid method in DUC 2005. In Proceedings of the Document Understanding Conference (DUC 05), Vancouver, B.C., Canada.

Ellen M. Voorhees and Donna Harman. 1999. Overview of the seventh text retrieval conference. In Proceedings of the Seventh Text Retrieval Conference (TREC-7). NIST Special Publication.