The Second PASCAL Recognising Textual Entailment Challenge

Roy Bar-Haim*, **Ido Dagan***, **Bill Dolan****, **Lisa Ferro**†, **Danilo Giampiccolo**‡,
Bernardo Magnini*, **Idan Szpektor***

*Computer Science Department, Bar-Ilan University, Ramat Gan 52900, Israel
**Microsoft Research, Redmond, WA 98052, USA
†The MITRE Corporation, 202 Burlington Rd., Bedford, MA 01730, USA
‡CELCT, Via dei Solteri 38, Trento, Italy
*ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica 38050 Povo, Trento, Italy

Abstract

This paper describes the Second PASCAL Recognising Textual Entailment Challenge (RTE-2).[1] We describe the RTE-2 dataset and overview the submissions for the challenge. One of the main goals for this year's dataset was to provide more "realistic" text-hypothesis examples, based mostly on outputs of actual systems. The 23 submissions for the challenge present diverse approaches and research directions, and the best results achieved this year are considerably higher than last year's state of the art.

1 Introduction

1.1 Textual entailment recognition

Textual entailment recognition is the task of deciding, given two text fragments, whether the meaning of one text is entailed (can be inferred) from another text (see section 2.2 for the specific operational definition of textual entailment assumed in the challenge). This task, introduced by Dagan and Glickman (2004), captures generically a broad range of inferences that are relevant for multiple applications. For example, a Question Answering (QA) system has to identify texts that entail the expected answer. Given the question *"Who is John Lennon's widow?"*, the text *"Yoko Ono unveiled a bronze statue of her late husband, John Lennon, to complete the official renaming of England's Liverpool Airport as Liverpool John Lennon Airport."* entails the expected answer *"Yoko Ono is John Lennon's widow"*. Similarly, semantic inference needs of other text understanding applications such as Information Retrieval (IR), Information Extraction (IE), and Machine Translation evaluation can be cast as entailment recognition (Dagan et al., 2006).

Textual entailment may serve as a unifying generic framework for modeling semantic inference, which so far have been addressed independently by separate application-specific research communities. Its formulation as a mapping between texts makes it independent of concrete semantic interpretations, which then become a mean rather than the goal. For example, in word sense disambiguation, it is not always easy to define explicitly the right *set of senses* to choose from. In practice, however, it is sufficient for most applications to determine whether a word meaning in a given context *lexically entails* another word. For instance, the occurrence of the word *"chair"* in the sentence *"IKEA announced a new comfort chair"* entails *"furniture"*, while its occurrence in the sentence *"MIT announced a new CS chair position"* does not. Thus, proper modeling of lexical entailment in context may alleviate the need to *interpret* word occurrences into explicitly stipulated senses.

Eventually, we hope that research on textual entailment will lead to the development of entailment "engines", which will 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).

[1] http://www.pascal-network.org/Challenges/RTE2

1.2 The First RTE Challenge

The first PASCAL Recognising Textual Entailment Challenge (RTE-1), (Dagan et al., 2006) introduced the first benchmark for the entailment recognition task. The RTE-1 dataset consists of manually collected text fragment pairs, termed *text (t)* (1-2 sentences) and *hypothesis (h)* (one sentence). The participating systems were required to judge for each pair whether *t* entails *h*. The pairs represented success and failure settings of inferences in various application types (termed *"tasks"*), including, among others, the QA, IE, IR, and MT evaluation described above.

The challenge raised noticeable attention in the research community, attracting 17 submissions from diverse groups. The relatively low accuracy achieved by the participating systems suggested that the entailment task is indeed a challenging one, with a wide room for improvement. It was followed by an ACL 2005 Workshop on Empirical Modeling of Semantic Equivalence and Entailment. The challenge and its dataset motivated further research on empirical entailment, which resulted in a number of publications in recent main conferences.[2]

1.3 Goals for the second challenge

Following the success and impact of RTE-1, the main goal of the second challenge was to support the continuation of research on textual entailment. Our main focus in creating the RTE-2 dataset was to provide more ”realistic” text-hypothesis examples, based mostly on outputs of actual systems. As in the previous challenge, the main task is judging whether a hypothesis *h* is entailed by a text *t*. The examples represent different levels of entailment reasoning, such as lexical, syntactic, morphological and logical. Data collection and annotation processes were improved this year, including cross-annotation of the examples across the organizing sites (most of the pairs were triply annotated). The data collection and annotation guidelines were revised and expanded. In order to make the challenge data more accessible, we also provided some pre-processing for the examples, including sentence splitting and dependency parsing.

2 The RTE-2 Dataset

2.1 Overview

The RTE-2 dataset consists of 1600 text-hypothesis pairs, divided into a development set and a test set, each containing 800 pairs. We followed the basic setting of RTE-1: the texts consist of 1-2 sentences, while the hypotheses are one sentence (usually shorter).

We chose to focus on four out of the seven applications that were present in RTE-1: Information Retrieval (IR), Information Extraction (IE), Question Answering (QA), and multi-document summarization (SUM[3]). Within each application setting the annotators selected positive entailment examples (annotated *YES*), where *t* does entail *h*, as well as negative examples (annotated *NO*), where entailment does not hold (50%-50% split, as in RTE-1). In total, 200 pairs were collected for each application in each dataset. Each pair was annotated with its related task (IE/IR/QA/SUM) and entailment judgment (YES/NO, released only in the development set). Some of the pairs in the development set are listed in Table 1.

The examples in the dataset are based mostly on outputs (both correct and incorrect) of Web-based systems, while most of the input was sampled from existing application-specific benchmarks[4]. Thus, the examples give some sense of how existing systems could benefit from an entailment engine post-processing their output. The data collection procedure for each task is described in sections 2.3 through 2.6.

2.2 Defining and judging entailment

We consider an applied notion of textual entailment, defined as a directional relation between two text fragments, termed *t* - the entailing text, and *h* - the entailed text. We say that *t* entails *h* if, typically, a human reading *t* would infer that *h* is most likely true. This somewhat informal definition is based on (and assumes) common human understanding of language as well as common background knowledge. Textual entailment recognition is the task of deciding, given *t* and *h*, whether *t* entails *h*.

[2] A list of related publications can be found at http://www.pascal-network.org/Challenges/RTE2/Introduction

[3] Equivalent to the CD task in RTE-1

[4] See the Acknowledgments section for a complete list of systems and benchmarks used

ID	Text	Hypothesis	Task	Judgment
77	Google and NASA announced a working agreement, Wednesday, that could result in the Internet giant building a complex of up to 1 million square feet on NASA-owned property, adjacent to Moffett Field, near Mountain View.	Google may build a campus on NASA property.	SUM	YES
110	Drew Walker, NHS Tayside's public health director, said: "It is important to stress that this is not a confirmed case of rabies."	A case of rabies was confirmed.	IR	NO
294	Meanwhile, in an exclusive interview with a TIME journalist, the first one-on-one session given to a Western print publication since his election as president of Iran earlier this year, Ahmadinejad attacked the "threat" to bring the issue of Iran's nuclear activity to the UN Security Council by the US, France, Britain and Germany.	Ahmadinejad is a citizen of Iran.	IE	YES
387	About two weeks before the trial started, I was in Shapiro's office in Century City.	Shapiro works in Century City.	QA	YES
415	The drugs that slow down or halt Alzheimer's disease work best the earlier you administer them.	Alzheimer's disease is treated using drugs.	IR	YES
691	Arabic, for example, is used densely across North Africa and from the Eastern Mediterranean to the Philippines, as the key language of the Arab world and the primary vehicle of Islam.	Arabic is the primary language of the Philippines.	QA	NO

Table 1: Examples of text-hypothesis pairs, taken from the RTE-2 development set

Some additional judgment criteria and guidelines are listed below (examples are taken from Table 1):

- Entailment is a directional relation. The hypothesis must be entailed from the given text, but the text need not be entailed from the hypothesis.

- The hypothesis must be fully entailed by the text. Judgment would 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) are judged as YES. For instance, in pair #387 one could claim that although Shapiro's office is in Century City, he actually never arrives to his office, and works elsewhere. However, this interpretation of *t* is very unlikely, and so the entailment holds with high probability. On the other hand, annotators were guided to avoid vague examples for which inference has some positive probability which is not clearly very high.

- Our definition of entailment allows presupposition of common knowledge, such as: a company has a CEO, a CEO is an employee of the company, an employee is a person, etc. For instance, in pair #294, the entailment depends on knowing that the president of a country is also a citizen of that country.

2.3 Collecting IE pairs

This task is inspired by the Information Extraction (and Relation Extraction) application, adapting the setting to pairs of texts rather than a text and a structured template. The pairs were generated using four different approaches. In the first approach, ACE-2004 relations (the relations tested in the ACE-2004 RDR task) were taken as templates for hypotheses. Relevant news articles were collected as texts (*t*). These collected articles were then given to actual IE systems for extraction of ACE relation instances. The system outputs were used as hypotheses, generating both positive examples (from correct outputs) and negative examples (from incorrect outputs). In the second approach, the output of IE systems on the dataset of the MUC-4 TST3 task (in which the events are acts of terrorism) was similarly used to create entailment pairs. In the third approach, additional entailment pairs were manually generated from both the annotated MUC-4 dataset and news articles collected for the ACE relations. For example, given the ACE relation "X work for Y" and the text "An Afghan interpreter, employed by the United States, was also wounded." (*t*), a hypothesis "An interpreter worked for Afghanistan." is created, producing a non-entailing (negative) pair. In the forth approach, hypotheses which correspond to new types of semantic relations (not found in the ACE and MUC datasets) were manually generated for sentences in collected news articles. These relations were taken from various semantic fields, such as sports, entertainment and science. All these processes 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.

2.4 Collecting IR pairs

In this application setting, the hypotheses are propositional IR queries, which specify some statement, e.g. "Alzheimer's disease is treated using drugs". The hypotheses were adapted and simplified from standard IR evaluation datasets (TREC and CLEF). Texts (*t*) that do or do not entail the hypothesis were selected from documents retrieved by different search engines (e.g. Google, Yahoo and MSN) for each hypothesis. In this application setting it is assumed that relevant documents (from an IR perspective) should entail the given propositional hypothesis.

2.5 Collecting QA pairs

Annotators were given questions, taken from TREC-QA and QA@CLEF datasets and the corresponding answers extracted from the Web by QA systems. Transforming a question-answer pair to text-hypothesis pair consisted of the following stages: First, the annotators picked from the answer passage an answer term of the expected answer type, either a correct or an incorrect one. Then, the annotators turned the question into an affirmative sentence with the answer term "plugged in". These affirmative sentences serve as the hypotheses (*h*), and the original answer passage serves as the text (*t*). For example (pair #575 in the development set), given the question *"How many inhabitants does Slovenia have?"* and an answer text *"In other words, with its 2 million inhabitants, Slovenia has only 5.5 thousand professional soldiers"* (*t*), the annotators picked *"2 million inhabitants"* as the (correct) answer term, which was used to turn the question into the statement *"Slovenia has 2 million inhabitants"* (*h*), producing a positive entailment pair. This process simulates the need of a QA system to verify that the retrieved passage text indeed entails the provided answer.

2.6 Collecting SUM pairs

In this setting *t* and *h* are sentences taken from a news document cluster, a collection of news articles that describe the same news item. Annotators were given output of multi-document summarization systems, including the document clusters and the summary generated for each cluster. The annotators 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 the same manner. This process simulates the need of a summarization system to identify information redundancy, which should be avoided in the summary, and may also increase the assessed importance of such repeated information.

2.7 Creating the final dataset

Cross-annotation of the collected pairs was done between the organizing sites. Each pair was judged by at least two annotators and most of the pairs (75% of the pairs in the development set, and all of the test set) were triply judged. As in RTE-1, we filtered out pairs on which the annotators disagreed. The average agreement on the test set (between each pair of annotators who shared at least 100 examples), was 89.2%, with average Kappa level of 0.78, which corresponds to “substantial agreement” (Landis and Koch, 1997). 18.2% of the pairs were removed from the test set due to disagreement. The following situations often caused disagreement:

- *t* gives approximate numbers and *h* gives exact numbers.
- *t* states an asserted claim made by some entity, and the *h* drops the assertion and just states the claim. For example:
  t: “Scientists say that global warming is made worse by human beings.”
  h: “Global warming is made worse by human beings.”
- *t* makes a weak statement, and *h* makes a slightly stronger statement about the same thing.

Additional filtering was done by two of the organizers, who discarded pairs that seemed controversial, too difficult, or redundant (too similar to other pairs). In this phase, 25.5% of the (original) pairs were removed from the test set.

We allowed only minimal correction of texts extracted from the web, e.g. fixing spelling and punctuation but not style, therefore the English of some of the pairs is less-than-perfect. In addition to the corrections made by the annotators, a final proofreading pass over the dataset was performed by one of the annotators.

3 The RTE-2 Challenge

3.1 Submission instructions and evaluation measures

The main task in the RTE-2 challenge was *classification* – entailment judgement for each pair in the test set. The evaluation criterion for this task was *accuracy* – the percentage of pairs correctly judged.

A secondary task was *ranking* the pairs, according to their entailment confidence. In this ranking, the first pair is the one for which the entailment is most certain, and the last pair is the one for which the entailment is least likely (i.e. the one for which the judgment as “NO” is the most certain). A perfect ranking would place all the positive pairs (for which the entailment holds) before all the negative pairs. This task was evaluated using the *Average precision* measure, which is a common evaluation measure for ranking (e.g. in information retrieval), 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 YES (Voorhees and Harman, 1999). More formally, it can be written as follows:

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

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 (note the difference between this measure and the Confidence Weighted Score used in the first challenge).

Participating teams were allowed to submit results of up to two systems, and many of the participants made use of this option, and provided the results of two runs.

3.2 Submitted systems

Twenty-three teams participated in the challenge, a 35% growth compared to last year. Table 2 lists for each submitted run the methods used and the obtained results. These methods include lexical overlapping, based on lexicons such as WordNet (Miller, 1995) and automatically acquired resources which are based on statistical measures; n-gram matching and subsequence overlapping be-

			System Description								
First Author (Group)	**Accuracy**	**Average Precision**	Lexical Relation DB	n-gram/ Subsequence overlap	Syntactic Matching/ Alignment	Semantic Role Labelling/ Framenet/ Propbank	Logical Inference	Corpus/ Web-Based Statistics	ML Classification	Paraphrase Templates/ Background Knowledge	Acquisition of Entailment Corpora
Adams (Dallas)	0.6262	0.6282	X						X		
Bos (Rome & Leeds)	0.6162	0.6689	X					X	X		
	0.6062	0.6042	X				X	X	X	X	
Burchardt (Saarland)	0.5900		X			X					
	0.5775		X		X	X			X		
Clarke (Sussex)	0.5275	0.5254		X				X			
	0.5475	0.5260		X							
de Marneffe (Stanford)	0.5763	0.6131	X		X			X	X	X	
	0.6050	0.5800	X		X			X	X	X	
Delmonte (Venice)	0.5475	0.5495	X		X	X				X	
Ferrández (Alicante)	0.5563	0.6089	X		X			X			
	0.5475	0.5743	X		X						
Herrera (UNED)	0.5975	0.5663	X						X		
	0.5887		X		X				X		
Hickl (LCC)	0.7538	0.8082	X	X	X	X		X	X		X
Inkpen (Ottawa)	0.5800	0.5751	X	X	X				X		
	0.5825	0.5816	X	X	X				X		
Katrenko (Amsterdam)	0.5900				X						
	0.5713										
Kouylekov (ITC-irst & Trento)	0.5725	0.5249	X		X			X			
	0.6050	0.5046	X		X			X	X		
Kozareva (Alicante)	0.5487	0.5589	X	X				X	X		
	0.5500	0.5485	X	X				X	X		
Litkowski (CL Research)	0.5813										
	0.5663				X						
Marsi (Tilburg & Twente)	0.6050		X		X			X			
Newman (Dublin)	0.5250	0.5052	X	X				X	X		
	0.5437	0.5103	X	X	X			X	X		
Nicholson (Melbourne)	0.5288	0.5464	X		X				X		
	0.5088	0.5053	X		X				X		
Nielsen (Colorado)	0.5962	0.6464		X	X			X	X		
	0.5875	0.6487		X	X			X	X		
Rus (Memphis)	0.5900	0.6047	X		X						
	0.5837	0.5785			X						
Schilder (Thomson & Minnesota)	0.5437		X		X			X	X		
	0.5550		X		X			X	X		
Tatu (LCC)	0.7375	0.7133	X				X			X	
Vanderwende (Microsoft Research & Stanford)	0.6025	0.6181	X		X			X	X		
	0.5850	0.6170	X		X			X			
Zanzotto (Milan & Rome)	0.6388	0.6441	X		X			X	X		
	0.6250	0.6317	X		X			X	X		

Table 2: Submission results and system description. Systems for which no component is indicated used lexical overlap.

tween *t* and *h*; syntactic matching, e.g. relation matching, and tree edit distance algorithms; semantic annotation induced using resources such as FrameNet (Baker et al., 1998); logical inference using logic provers; statistics computed from local corpora or the Web (including statical measures available for lexical resources such as WordNet); usage of background knowledge, including inference rules and paraphrase templates, and acquisition (automatic and manual) of additional entailment corpora. Many of the systems derive multiple similarity measures, based on different levels of analysis (lexical, syntactic, logical), and subsequently use them as features for a classifier that makes the final decision.

Overall, the common criteria for entailment recognition were *similarity* between *t* and *h*, or the *coverage* of *h* by *t* (in lexical and lexical syntactic methods), and the ability to *infer h* from *t* (in the logical approach). Zanzotto et al. also measured the similarity between different *(t,h) pairs* (cross-pair similarity). Some groups tried to detect non-entailment, by looking for various kinds of mismatch between the text and the hypothesis. This approach is related to an earlier observation in (Vanderwende et al., 2005), which suggested that it is often easier to detect false entailment.

3.3 Results

The accuracy achieved by the participating systems ranges from 53% to 75% (considering the best run of each group), while most of the systems obtained 55%-61%. Two submissions, Hickl at el. (accuracy 75.4%, average precision 80.8%) and Tatu at el. (accuracy 73.8%, average precision 71.3%), stand out as about 10% higher than the other systems. The top accuracies are considerably higher than the best results achieved in RTE-1 (around 60%).

The results show, for the first time, that systems that rely on deep analysis such as syntactic matching and logical inference can considerably outperform lexical systems, which were shown to achieve around 60% on the RTE datasets. In the RTE-1 challenge, one of the two best performing systems was based on lexical statistics from the web (Glickman et al., 2006). Zanzotto at el. experimented with baseline lexical systems, applied to both RTE-1 and RTE-2 datasets. For RTE-1 they found that even a simple statistical lexical system, based on IDF measure, gets close to 60% in accuracy. Bar-Haim et al. (Bar-Haim et al., 2005) also showed by manually analyzing the RTE-1 dataset, that lexical systems are expected to achieve up to around 60%, if we require that *h* is fully lexically entailed from (covered by) *t*. For the RTE-2 test set, Zanzotto et al. found that simple lexical overlapping achieves accuracy of 60%, better than any other sophisticated lexical methods they tested (Katrenko and Adriaans report 57% for a slightly different baseline).

3.4 The contribution of knowledge and training data

Although it is clear that deeper analysis is a must for achieving high accuracy, most of the systems participated in RTE-2 that employed deep analysis did not improve significantly over the 60% baseline of lexical matching. The participants' reports point out two main reasons for the shortcoming of current systems: the size of the training corpus (RTE-2 development set and the RTE-1 datasets together contain less than 2,200 pairs), and the lack of linguistic and background knowledge.

It seems that the best performing systems were those which better coped with these issues. Hickl et al. utilized a very large entailment corpus, automatically collected from the web, following (Burger and Ferro, 2005). In addition, they manually annotated a corpus of lexical entailment, which was used to bootstrap automatic annotation of a larger lexical entailment corpus. These corpora contributed 10% to the overall accuracy they achieved. Tatu et al. developed an entailment system based on logical inference, which relys on extensive linguistic and background knowledge from various sources.

The success of these systems suggests that perhaps the most important factors for deep entailment systems are the amount of linguistic and background knowledge, and the size of training corpora, rather than the exact method for modeling *t* and *h* and the exact inference mechanism.

3.5 Per-task analysis

Per-task analysis shows that systems scored considerably higher on the multi-document summarization task (SUM). The same trend was observed in RTE-1 for the comparable documents (CD) task, which was similar to the RTE-2 summarization task. For most systems, the lowest accuracy was obtained for the IE task. Katrenko and Adriaans report that simple lexical overlapping was able to predict correctly entailment for 67% of the SUM pairs, but only for 47% of the IE pairs.

Some of the participants took into account such inter-task differences, and tuned the parameters of their models separately for each task. Given the observed differences among the tasks, it seems that better understanding of how entailment in each task differs might improve the performance of future systems.

3.6 Additional observations

Some participants tested their systems on both RTE-1 and RTE-2 datasets. Some systems performed better on RTE-1 while others performed better on RTE-2, and the results were usually quite close, with up to 5% difference for either side. This indicates similar level of difficulty for both datasets. However, simple lexical overlap systems were found to perform better on the RTE-2 test set than on RTE-1 test set - 60% on RTE-2 vs. 53.9% on RTE-1, as reported by Zanzotto et al., (although for the RTE-1 development set they obtained 57.1%). Interestingly, de Marneffe et al. and Zanzotto et al. report that adding the RTE-1 data to the RTE-2 training set reduced the results, which indicates the variance between the two datasets (notice that the RTE-1 datasets include three tasks not present in RTE-2. Inkpen at el. showed that the results somewhat improve if only the compatible tasks in RTE-1 are considered). Schilder and Thomson McInnes found that classification using only the lengths of *t* and *h* as features could give accuracy of 57.4%.

In the RTE-2 dataset (both the development set and the test set), multiple IR pairs were created for a single IR query (where *t* was extracted from different documents retrieved), and similarly, multiple QA pairs were created for a single question (where *t* was extracted from different answer passages). Some of the groups (de Marneffe et al., Nicholson et al.) noted that these dependencies between the pairs could potentially have a negative effect on the learning, and somewhat bias the evaluation on the test set. In practice, however, there was no evidence that systems perform significantly worse on the RTE-2 test set than on the RTE-1 test set (using RTE-2/RTE-1 development sets, respectively, for training), and, as described above, similar scores were obtained for both datasets.

4 Conclusion and future work

The submissions for the Second PASCAL Recognising Textual Entailment Challenge show growing interest in this applied framework. The considerable improvement in performance achieved within only one year is very encouraging, and the diversity of new approaches and research directions introduced this year seem very promising for further research. While the setting for the entailment recognition task in RTE-2 followed the same setting of RTE-1, we expect that the next RTE challenges will introduce new settings. One possible direction is to provide wider contexts, i.e. expanding *t* from 1-2 sentences to paragraphs or even complete documents.

Acknowledgment

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

- AnswerBus question answering system, provided by Zhiping Zheng, Computational Linguistics Department, Saarland University.
  http://answerbus.coli.uni-sb.de/

- PowerAnswer question answering system, from Language Computer Corporation, provided by Dan Moldovan, Abraham Fowler, Christine Clark, Arthur Dexter and Justin Larrabee.
  http://www.languagecomputer.com/solutions/question_answering/power_answer/

- 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/

- IBM's information extraction system, provided by Salim Roukos and Nanda Kambhatla, I.B.M. T.J. Watson Research Center.

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

- ITC-irst's information extraction system, provided by Lorenza Romano, Cognitive and Communication Technologies (TCC) division, ITC-irst, Trento, Italy.

- 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/

We would like to thank the people and organizations that made these sources available for the challenge. In addition, we thank Oren Glickman and Dan Roth for their assistance and advice.

We would also like to acknowledge the people and organizations involved in creating and annotating the data: Malky Rabinowitz, Dana Mills, Ruthie Mandel, Errol Hayman, Vanessa Sandrini, Allesandro Valin, Elizabeth Lima, Jeff Stevenson, Amy Muia and the Butler Hill Group.

This work was supported in part by the IST Programme of the European Community, under the *PASCAL Network of Excellence*, IST-2002-506778. This publication only reflects the authors' views. We wish to thank the managers of the PASCAL challenges program, Michele Sebag, Florence d'Alche-Buc, and Steve Gunn, and the PASCAL Challenges Workshop Chair, Rodolfo Delmonte, for their efforts and support, which made this challenge possible.

References

C. Baker, C. Fillmore, and J. Lowe. 1998. The Berkeley Framenet project. In *Proceedings of the COLING-ACL*, Montreal, Canada.

Roy Bar-Haim, Idan Szpecktor, and Oren Glickman. 2005. Definition and analysis of intermediate entailment levels. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 55–60, Ann Arbor, Michigan, June. Association for Computational Linguistics.

John Burger and Lisa Ferro. 2005. Generating an entailment corpus from news headlines. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 49–54, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. PASCAL workshop on Text Understanding and Mining.

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

Oren Glickman, Ido Dagan, and Moshe Koppel. 2006. Web based probabilistic textual entailment. In Quiñonero-Candela et al., editor, *MLCW 2005, LNAI Volume 3944*, pages 287–298. Springer-Verlag.

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

G. A. Miller. 1995. WordNet: A Lexical Databases for English. *Communications of the ACM*, pages 39–41, November.

Lucy Vanderwende, Deborah Coughlin, and Bill Dolan. 2005. What syntax can contribute in entailment task. Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment (and forthcoming LNAI book chapter).

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.