The PASCAL Recognising Textual Entailment Challenge

Ido Dagan[1], Oren Glickman[1], and Bernardo Magnini[2]

[1] Bar Ilan University, Ramat Gan 52900, Israel
{dagan, glikmao}@cs.biu.ac.il
http://cs.biu.ac.il/~{dagan, glikmao}/
[2] ITC-irst, 38100 Trento, Italy
magnini@itc.it
http://tcc.itc.it/people/magnini.html

Abstract. This paper describes the PASCAL Network of Excellence first *Recognising Textual Entailment* (RTE-1) Challenge benchmark[1]. The RTE task is defined as recognizing, given two text fragments, whether the meaning of one text can be inferred (entailed) from the other. This application-independent task is suggested as capturing major inferences about the variability of semantic expression which are commonly needed across multiple applications. The Challenge has raised noticeable attention in the research community, attracting 17 submissions from diverse groups, suggesting the generic relevance of the task.

1 Introduction

1.1 Rational

A fundamental phenomenon of natural language is the variability of semantic expression, where the same meaning can be expressed by, or inferred from, different texts. This phenomenon may be considered as the dual problem of language ambiguity, together forming the many-to-many mapping between language expressions and meanings. Many natural language processing applications, such as Question Answering (QA), Information Extraction (IE), (multi-document) summarization, and machine translation (MT) evaluation, need a model for this variability phenomenon in order to recognize that a particular target meaning can be inferred from different text variants.

Even though different applications need similar models for semantic variability, the problem is often addressed in an application-oriented manner and methods are evaluated by their impact on final application performance. Consequently it becomes difficult to compare, under a generic evaluation framework, practical inference methods that were developed within different applications. Furthermore, researchers within one application area might not be aware of relevant methods that were developed in the context of another application. Overall,

[1] See http://www.pascal-network.org/Challenges/RTE/ for the first and second RTE challenges.

J. Quiñonero-Candela et al. (Eds.): MLCW 2005, LNAI 3944, pp. 177–190, 2006.

there seems to be a lack of a clear framework of generic task definitions and evaluations for such "applied" semantic inference, which also hampers the formation of a coherent community that addresses these problems. This situation might be confronted, for example, with the state of affairs in syntactic processing, where clear application-independent tasks, communities (and even standard conference session names) have matured.

The *Recognising Textual Entailment* (RTE) Challenge is an attempt to promote an abstract generic task that captures major semantic inference needs across applications. The task requires to recognize, given two text fragments, whether the meaning of one text can be inferred (entailed) from another text. More concretely, our applied notion of *textual entailment* is defined as a directional relationship between pairs of text expressions, denoted by T - the entailing "Text", and H - the entailed "Hypothesis". 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. It is similar in spirit to evaluation of applied tasks such as Question Answering and Information Extraction, in which humans need to judge whether the target answer or relation can indeed be inferred from a given candidate text. Table 1 includes a few examples from the dataset along with their gold standard annotation.

As in other evaluation tasks our definition of textual entailment is operational, and corresponds to the judgment criteria given to the annotators who decide whether this relationship holds between a given pair of texts or not. Recently there have been just a few suggestions in the literature to regard entailment recognition for texts as an applied, empirically evaluated, task (see [4], [6] and [12]).

It seems that major inferences, as needed by multiple applications, can indeed be cast in terms of textual entailment. For example, a QA system has to identify texts that entail a hypothesized answer. Given the question "*Who painted 'The Scream'?*", the text "*Norway's most famous painting, 'The Scream' by Edvard Munch,...*" entails the hypothesized answer form "Edvard Munch painted 'The Scream'." (see corresponding example 568 in Table 1). Similarly, for certain Information Retrieval queries the combination of semantic concepts and relations denoted by the query should be entailed from relevant retrieved documents. In IE entailment holds between different text variants that express the same target relation. In multi-document summarization a redundant sentence, to be omitted from the summary, should be entailed from other sentences in the summary. And in MT evaluation a correct translation should be semantically equivalent to the gold standard translation, and thus both translations should entail each other. Consequently, we hypothesize that textual entailment recognition is a suitable generic task for evaluating and comparing applied semantic inference models. Eventually, such efforts can promote the development of entailment recognition "engines" which may provide useful generic modules across applications.

Our applied notion of Textual entailment is also related, of course, to classical semantic entailment in the linguistics literature. A common definition of

Table 1. Examples of Text-Hypothesis pairs

ID	TEXT	HYPOTHESIS	TASK	VALUE
568	*Norway's most famous painting, "The Scream" by Edvard Munch, was recovered Saturday, almost three months after it was stolen from an Oslo museum.*	*Edvard Munch painted "The Scream".*	QA	True
1586	*The Republic of Yemen is an Arab, Islamic and independent sovereign state whose integrity is inviolable, and no part of which may be ceded.*	*The national language of Yemen is Arabic.*	QA	True
1076	*Most Americans are familiar with the Food Guide Pyramid– but a lot of people don't understand how to use it and the government claims that the proof is that two out of three Americans are fat.*	*Two out of three Americans are fat.*	RC	True
1667	*Regan attended a ceremony in Washington to commemorate the landings in Normandy.*	*Washington is located in Normandy.*	IE	False
13	*iTunes software has seen strong sales in Europe.*	*Strong sales for iTunes in Europe.*	IR	True
2016	*Google files for its long awaited IPO.*	*Google goes public.*	IR	True
2097	*The economy created 228,000 new jobs after a disappointing 112,000 in June.*	*The economy created 228,000 jobs after dissapointing the 112,000 of June.*	MT	False
893	*The first settlements on the site of Jakarta were established at the mouth of the Ciliwung, perhaps as early as the 5th century AD.*	*The first settlements on the site of Jakarta were established as early as the 5th century AD.*	CD	True
1960	*Bush returned to the White House late Saturday while his running mate was off campaigning in the West.*	*Bush left the White House.*	PP	False
586	*The two suspects belong to the 30th Street gang, which became embroiled in one of the most notorious recent crimes in Mexico: a shootout at the Guadalajara airport in May, 1993, that killed Cardinal Juan Jesus Posadas Ocampo and six others.*	*Cardinal Juan Jesus Posadas Ocampo died in 1993.*	QA	True
908	*Time Warner is the world's largest media and Internet company.*	*Time Warner is the world's largest company.*	RC	False
1911	*The SPD got just 21.5% of the vote in the European Parliament elections, while the conservative opposition parties polled 44.5%.*	*The SPD is defeated by the opposition parties.*	IE	True

entailment in formal semantics ([3]) specifies that a text t entails another text h (hypothesis, in our terminology) if h is true in every circumstance (*possible world*) in which t is true. For example, in example 13 from Table 1 we'd assume humans to agree that the hypothesis is necessarily true in any circumstance for which the text is true. In such intuitive cases, our proposed notion of textual entailment corresponds to the classical notions of semantic entailment.

However, our applied definition allows for cases in which the truth of the hypothesis is highly plausible, for most practical purposes, rather than certain. In Table 1, examples 1586, 1076, 893 and 586 were annotated as True even though the entailment in this cases is not certain. This seems to match the types of uncertain inferences that are typically expected from text based applications. [7] present a first attempt to define in probabilistic terms a coherent notion and generative setting of textual entailment. For a discussion on the relation between Textual Entailment and some classical linguistic notions such as presupposition and implicature see [16]. There is also considerable classical work on fuzzy or uncertain inference (e.g. [1], [8], [9]). Making significant reference to this rich body of literature and deeply understanding the relationships between our operational textual entailment definition and relevant linguistic notions is an ongoing research topic, and is beyond the scope of this paper. Finally, it may be noted that from an applied empirical perspective much of the effort is directed at recognizing meaning-entailing variability at rather shallow linguistic levels, rather than addressing relatively delicate logical issues as typical in classical literature.

1.2 The Challenge Scope

As a first step towards the above goal we created a dataset of Text-Hypothesis (T-H) pairs of small text snippets, corresponding to the general news domain (see Table 1). Examples were manually labeled for entailment - whether T entails H or not - by human annotators, and were divided into *development* and *test* datasets. Participating systems were asked to decide for each T-H pair whether T indeed entails H (denoted as *True*) or not (*False*), and results were compared to the manual gold standard.

The dataset was collected with respect to different text processing applications, as detailed in the next section. Each portion of the dataset was intended to include typical T-H examples that may correspond to success and failure cases of the actual applications. The collected examples represent a range of different levels of entailment reasoning, based on lexical, syntactic, logical and world knowledge, at different levels of difficulty.

The distribution of examples in this challenge has been somewhat biased to choosing nontrivial pairs, and also imposed a balance of True and False examples. For this reason, systems performances in applicative settings might be different than the figures for the challenge data, due to different distributions of examples in particular applications. Yet, the data does challenge systems to handle properly a broad range of entailment phenomena. Overall, we were aiming at an explorative rather than a competitive setting, hoping that meaningful baselines and analyses for the capabilities of current systems will be obtained.

Finally, the task definition and evaluation methodologies are clearly not mature yet. We expect them to change over time and hope that participants' contributions, observations and comments will help shaping this evolving research direction.

2 Dataset Preparation and Application Settings

The dataset of Text-Hypothesis pairs was collected by human annotators. It consists of seven subsets, which correspond to typical success and failure settings in different applications, as listed below. Within each application setting the annotators selected both positive entailment examples (*True*), where T is judged to entail H, as well as negative examples (False), where entailment does not hold (a 50%-50% split). Typically, T consists of one sentence (sometimes two) while H was often made a shorter sentence (see Table 1). The full datasets are available for download at the Challenge website[2].

In some cases the examples were collected using external sources, such as available datasets or systems (see Acknowledgements), while in other cases examples were collected from the web, focusing on the general news domain. In all cases the decision as to which example pairs to include was made by the annotators. The annotators were guided to obtain a reasonable balance of different types of entailment phenomena and of levels of difficulty. Since many T-H pairs tend to be quite difficult to recognize, the annotators were biased to limit the proportion of difficult cases, but on the other hand to try avoiding high correlation between entailment and simple word overlap. Thus, the examples do represent a useful broad range of naturally occurring entailment factors. Yet, we cannot say that they correspond to a particular representative distribution of these factors, or of True vs. False cases, whatever such distributions might be in different settings. Thus, results on this dataset may provide useful indications of system capabilities to address various aspects of entailment, but do not predict directly the performance figures within a particular application.

It is interesting to note in retrospect that the annotators' selection policy yielded more negative examples than positive ones in the cases where T and H have a very high degree of lexical overlap. This anomaly was noticed also by Bos and Markert, Bayer et al. and Glickman et al. (this Volume), and affected the design or performance of their systems

2.1 Application Settings

Information Retrieval (IR). Annotators generated hypotheses (H) that may correspond to meaningful IR queries that express some concrete semantic relations. These queries are typically longer and more specific than a standard keyword query, and may be considered as representing a semantic-oriented variant within IR. The queries were selected by examining prominent sentences in news stories, and were then submitted to a web search engine. Candidate texts (T) were selected from the search engine's retrieved documents, picking candidate texts that either do or do not entail the hypothesis.

Comparable Documents (CD). Annotators identified T-H pairs by examining a cluster of comparable news articles that cover a common story. They

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

examined “aligned” sentence pairs that overlap lexically, in which semantic entailment may or may not hold. Some pairs were identified on the web using Google news[3] and others taken from an available resource of aligned English sentences (see Acknowledgments). The motivation for this setting is the common use of lexical overlap as a hint for semantic overlap in comparable documents, e.g. for multi-document summarization.

Reading Comprehension (RC). This task corresponds to a typical reading comprehension exercise in human language teaching, where students are asked to judge whether a particular assertion can be inferred from a given text story. The challenge annotators were asked to create such hypotheses relative to texts taken from news stories, considering a reading comprehension test for high school students.

Question Answering (QA). Annotators used the TextMap Web Based Question Answering system available online (see Acknowledgments). The annotators used a resource of questions from CLEF-QA[4] (mostly) and TREC[5], but could also construct their own questions. For a given question, the annotators chose first a relevant text snippet (T) that was suggested by the QA system as including the correct answer. They then turned the question into an affirmative sentence with the hypothesized answer ”plugged in” to form the hypothesis (H). For example, given the question, “Who is Ariel Sharon?” and taking a candidate answer text “Israel’s Prime Minister, Ariel Sharon, visited Prague” (T), the hypothesis H is formed by turning the question into the statement “Ariel Sharon is Israel’s Prime Minister”, producing a True entailment pair.

Information Extraction (IE). This task is inspired by the Information Extraction application, adapting the setting for pairs of texts rather than a text and a structured template. For this task the annotators used an available dataset annotated for the IE relations “kill” and “birth place” produced by UIUC (see acknowledgments), as well as general news stories in which they identified manually “typical” IE relations. Given an IE relation of interest (e.g. a purchasing event), annotators identified as the text (T) candidate news story sentences in which the relation is suspected to hold. As a hypothesis they created a straightforward natural language formulation of the IE relation, which expresses the target relation with the particular slot variable instantiations found in the text. For example, given the information extraction task of identifying killings of civilians, and a text “Guerrillas killed a peasant in the city of Flores.”, a hypothesis “Guerrillas killed a civilian” is created, producing a True entailment pair.

Machine Translation (MT). Two translations of the same text, an automatic translation and a gold standard human translation (see Acknowledgements), were compared and modified in order to obtain T-H pairs. The automatic translation

[3] http://news.google.com

[4] http://clef-qa.itc.it/

[5] http://trec.nist.gov/data/qa.html

was alternately taken as either T or H, where a correct translation corresponds to True entailment. The automatic translations were sometimes grammatically adjusted, being otherwise grammatically unacceptable.

Paraphrase Acquisition (PP). Paraphrase acquisition systems attempt to acquire pairs (or sets) of lexical-syntactic expressions that convey largely equivalent or entailing meanings. Annotators selected a text T from some news story which includes a certain relation, for which a paraphrase rule from a paraphrase acquisition system (see Acknowledgements) may apply. The result of applying the paraphrase rule on T was chosen as the hypothesis H. Correct paraphrases suggested by the system, which were applied in an appropriate context, yielded True T-H pairs; otherwise a False example was generated. For example, given the sentence "*The girl was found in Drummondville.*" and by applying the paraphrase rule $X\ was found in\ Y \Rightarrow Y\ contains\ X$, we obtain the hypothesis "*Drummondville contains the girl.*" Yielding a False example.

2.2 Additional Guidelines

Some additional annotation criteria and guidelines are listed below:

- Given that the text and hypothesis might originate from documents at different points in time, tense aspects are ignored.
- In principle, the hypothesis must be fully entailed by the text. Judgment would be False if the hypothesis includes parts that cannot be inferred from the text. However, cases in which inference is very probable (but not completely certain) are still judged at True. In example #586 in Table 1 one could claim that the shooting took place in 1993 and that (theoretically) the cardinal could have been just severely wounded in the shooting and has consequently died a few months later in 1994. However, this example is tagged as True since the context seems to imply that he actually died in 1993. To reduce the risk of unclear cases, annotators were guided to avoid vague examples for which inference has some positive probability that is not clearly very high.
- To keep the contexts in T and H self-contained annotators replaced anaphors with the appropriate reference from preceding sentences where applicable. They also often shortened the hypotheses, and sometimes the texts, to reduce complexity.
- Annotators were directed to assume common background knowledge of the news domain such as that a company has a CEO, a CEO is an employee of the company, an employee is a person, etc. However, it was considered unacceptable to presume highly specific knowledge, such as that Yahoo bought Overture for 1.63 billion dollars.

2.3 The Annotation Process

Each example T-H pair was first judged as True/False by the annotator that created the example. The examples were then cross-evaluated by a second judge, who received only the text and hypothesis pair, without any additional information from the original context. The annotators agreed in their judgment for roughly 80% of the examples, which corresponded to a 0.6 Kappa level (moderate agreement). The 20% of the pairs for which there was disagreement among the judges were discarded from the dataset. Furthermore, one of the organizers performed a light review of the remaining examples and eliminated an additional 13% of the original examples, which might have seemed controversial. Altogether, about 33% of the originally created examples were filtered out in this process.

The remaining examples were considered as the gold standard for evaluation, split to 567 examples in the development set and 800 in the test set, and evenly split to True/False examples. Our conservative selection policy aimed to create a dataset with non-controversial judgments, which will be addressed consensually by different groups. It is interesting to note that few participants have independently judged portions of the dataset and reached high agreement levels with the gold standard judgments, of 95% on all the test set (Bos and Markert), 96% on a subset of roughly a third of the test set (Vanderwende et al.) and 91% on a sample of roughly 1/8 of the development set (Bayer et al.).

3 Submissions and Results

3.1 Submission Guidelines

Submitted systems were asked to tag each *T*-*H* pair as either True, predicting that entailment does hold for the pair, or as False otherwise. In addition, systems could optionally add a confidence score (between 0 and 1) where 0 means that the system has no confidence of the correctness of its judgment, and 1 corresponds to maximal confidence. Participating teams were allowed to submit results of up to 2 systems or runs.

The development data set was intended for any system tuning needed. It was acceptable to run automatic knowledge acquisition methods (such as synonym collection) specifically for the lexical and syntactic constructs present in the test set, as long as the methodology and procedures are general and not tuned specifically for the test data[6].

In order to encourage systems and methods which do not cover all entailment phenomena we allowed submission of partial coverage results, for only part of the test examples. Naturally, the decision as to on which examples the system abstains were to be done automatically by the system (with no manual involvement).

3.2 Evaluation Criteria

The judgments (classifications) produced by the systems were compared to the gold standard. The percentage of matching judgments provides the accuracy of the run, i.e. the fraction of correct responses.

[6] We presumed that participants complied with this constraint. It was not enforced in any way.

As a second measure, a Confidence-Weighted Score (cws, also known as Average Precision) was computed. Judgments of the test examples were sorted by their confidence (in decreasing order), calculating the following measure:

$$cws = \frac{1}{n} \sum_{i=1}^{n} \frac{\#correct - up - to - rank - i}{i}$$

where n is the number of the pairs in the test set, and i ranges over the sorted pairs. The Confidence-Weighted Score ranges between 0 (no correct judgments at all) and 1 (perfect classification), and rewards the systems' ability to assign a higher confidence score to the correct judgments than to the wrong ones. Note that in the calculation of the confidence weighted score correctness is with respect to classification - i.e. a negative example, in which entailment does not hold, can be correctly classified as false. This is slightly different from the common use of average precision measures in IR and QA, in which systems rank the results by confidence of positive classification and correspondingly only true positives are considered correct.

3.3 Submitted Systems and Results

Sixteen groups submitted the results of their systems for the challenge data, while one additional group submitted the results of a manual analysis of the dataset (Vanderwende et al., see below). As expected, the submitted systems incorporated a broad range of inferences that address various levels of textual entailment phenomena. Table 2 presents some common (crude) types of inference components which, according to our understanding, were included in the various systems (see [2] and [13] who propose related breakdowns of inference types).

The most basic type of inference measures the degree of word overlap between T and H, possibly including stemming, lemmatization, part of speech tagging, and applying a statistical word weighting such as idf. Interestingly, a non-participating system that operated solely at this level, using a simple decision tree trained on the development set, obtained an accuracy level of 58%, which might reflect a knowledge-poor baseline (see [5]). Higher levels of lexical inference considered relationships between words that may reflect entailment, based either on statistical methods or WordNet. Next, some systems measured the degree of match between the syntactic structures of T and H, based on some distance criteria. Finally, few systems incorporated some form of "world knowledge", and a few more applied a logical prover for making the entailment inference, typically over semantically enriched representations. Different decision mechanisms were applied over the above types of knowledge, including probabilistic models, probabilistic Machine Translation models, supervised learning methods, logical inference and various specific scoring mechanisms.

Table 2 shows the results for the runs as submitted to the challenge (later post-submission results may appear in this Volume). Overall system accuracies were between 50 and 60 percent and system cws scores were between 0.50 and 0.70. Since the dataset was balanced in terms of true and false examples, a system

Table 2. Accuracy and cws results for the system submissions, ordered by first author. Partial coverage refers to the percentage of examples classified by the system out of the 800 test examples. (The results of the manual analysis by Vanderwende at al. (MSR) are summarized separately in the text.)

First Author (Group)	accuracy	cws	partial coverage	System description					
				Word overlap	Statistical lexical relations	WordNet	Syntactic matching	world knowledge	Logical inference
Akhmatova (Macquarie)	0.519	0.507		X					X
Andreevskaia (Concordia)	0.519	0.515				X	X		
	0.516	0.52							
Bayer (MITRE)	0.586	0.617			X				
	0.516	0.503	73%					X	X
Bos (Edinburgh & Leeds)	0.563	0.593		X		X		X	X
	0.555	0.586		X					
Delmonte (Venice & irst)	0.606	0.664	62%			X	X		X
Fowler (LCC)	0.551	0.56				X		X	X
Glickman (Bar Ilan)	0.586	0.572			X				
	0.53	0.535							
Herrera (UNED)	0.566	0.575		X	X		X		
	0.558	0.571		X					
Jijkoun (Amsterdam)	0.552	0.559		X	X				
	0.536	0.553		X		X			
Kouylekov (irst)	0.559	0.607		X	X		X		
	0.559	0.585							
Newman (Dublin)	0.563	0.592		X	X				
	0.565	0.6							
Perez (Madrid)	0.495	0.517		X					
	0.7	0.782	19%						
Punyakanok (UIUC)	0.561	0.569					X		
Raina (Stanford)	0.563	0.621			X	X	X		X
	0.552	0.686							
Wu (HKUST)	0.512	0.55			X		X		
	0.505	0.536							
Zanzotto (Rome-Milan)	0.524	0.557				X	X		
	0.518	0.559							

that uniformly predicts True (or False) would achieve an accuracy of 50% which constitutes a natural baseline. Another baseline is obtained by considering the distribution of results in random runs that predict True or False at random. A run with $cws > 0.540$ or $accuracy > 0.535$ is better than chance at the 0.05 level and a run with $cws > 0.558$ or $accuracy > 0.546$ is better than chance at the 0.01 level.

Unlike other system submissions, Vanderwende et al. (this Volume) report an interesting manual analysis of the test examples. Each example was analyzed as whether it could be classified correctly (as either True or False) by taking into account only syntactic considerations, optionally augmented by a lexical thesaurus. An "ideal" decision mechanism that is based solely on these levels of inference was assumed. Their analysis shows that 37% of the examples could (in principle) be handled by considering syntax alone, and 49% if a thesaurus is also consulted.

The Comparable Documents (CD) task stands out when observing the performance of the various systems broken down by tasks. Generally the results on this task are significantly higher than results on the other tasks with results as high as 87% accuracy and cws of 0.95. This behavior might indicate that in comparable documents there is a high prior probability that seemingly matching sentences indeed convey the same meanings. We also note that for some systems it is the success on this task which pulled the figures up from the insignificance baselines.

Our evaluation measures do not favor specifically recognition of positive entailment. A system which does well in recognizing when entailment does not hold would do just as well in terms of accuracy and cws as a system tailored to recognize true examples. In retrospect, standard measures of precision, recall and f in terms of the positive (entailing) examples would be appropriate as additional measures for this evaluation. In fact, some systems recognized only very few positive entailments (a recall between 10-30 percent). None of the systems performed significantly better than the f=0.67 baseline of a system which uniformly predicts true.

4 Discussion

As a new task and a first challenge, Textual Entailment Recognition is still making its first steps towards becoming a mature discipline within the Natural Language Processing community. We received a lot of feedback from the participants and other members of the research community, which partly contributed to the design of the second challenge (RTE-2) which is planned for 2006. Following are some issues that came up at the panels and discussions at the challenge workshop.

Multi Valued Annotation. In our setting we used a binary {True, False} annotation - a hypothesis is either entailed from the text or not. An annotation of False was used to denote both cases in which the truth value of the hypothesis is either (most likely) false or unknown given the text. Yet, one might want to distinguish between cases (such as example 1667 in Table 1) for which the hypothesis is False given the text and cases (such as example 2097) for which it is unknown whether the hypothesis is True or False. For this reason, a 3-valued annotation scheme ({True, False, Unknown}; see [10]) was proposed as a possible alternative. Furthermore, given the fuzzy nature of the task, it is not clear whether a 3-valued annotation would suffice and so n-valued annotation or even a Fuzzy logic scheme ([15]) may be considered as well. Allowing for a richer annotation scheme may enable to include the currently discarded examples on which there was no agreement amongst the annotators (see Section 2.3).

Assumed Background Knowledge. Textual inferences are based on information that is explicitly asserted in the text and often on additional assumed background knowledge not explicitly stated in the text. In our guidelines (see Section 2.2) we allowed annotators to assume common knowledge of the news domain. However, it is not clear how to separate out linguistic knowledge from world knowledge, and different annotators might not agree on what constitutes common background knowledge. For example, in example 1586 in Table 1 one needs to assume world knowledge regarding Arab states and the Arab language in order to infer the correctness of the hypothesis from the text. Furthermore, the criteria defining what constitutes acceptable background knowledge may be hypothesis dependant. For example, it is inappropriate to assume as background knowledge that The national language of Yemen is Arabic when judging example 1586, since this is exactly the hypothesis in question. On the other hand, such background knowledge might be assumed when examining the entailment “Grew up in Yemen” → “Speaks Arabic”. Overall, there seemed to be a consensus that it is necessary to assume the availability of background knowledge for judging entailment, even though it becomes one of the sources for certain disagreements amongst human annotators.

Common Preprocessing. Textual Entailment systems typically rely on the output of several NLP components prior to performing their inference, such as tokenization, lemmatization, part-of-speech tagging, named entity recognition and syntactic parsing. Since different systems differ in their preprocessing modules it becomes more difficult to compare them. In the next Challenge we plan to supply some common pre-processing of the data in order to enable better system comparison and to let participants focus on the inference components.

Entailment Subtasks. Textual entailment recognition is a complex task and systems typically perform multiple sub-tasks. It would therefore be interesting to define and compare performance on specific relevant subtasks. For example, [2] and [7] define lexical and lexical-syntactic entailment subtasks and [11] define an entailment-alignment subtask. Datasets that are annotated for such subtasks may be created in the future.

Inference Scope. Textual Entailment systems need to deal with a wide range of inference types. So far we were interested in rather direct inferences that are based mostly on information in the text and background knowledge. Specialized types of inference, such as temporal reasoning, complex logical inference or arithmetic calculations (see example 1911 from Table 1) were typically avoided but may be considered more systematically in the future.

5 Conclusions

The PASCAL *Recognising Textual Entailment* (RTE) Challenge is an initial attempt to form a generic empirical task that captures major semantic inferences across applications. The high level of interest in the challenge, demonstrated by the submissions from 17 diverse groups and noticeable interest in the research community, suggest that textual entailment indeed captures highly relevant core tasks.

The results obtained by the participating systems may be viewed as typical for a new and relatively difficult task (cf. for example the history of MUC benchmarks). Overall performance figures for the better systems were significantly higher than some baselines. Yet, the absolute numbers are relatively low, with small, though significant, differences between systems. Interestingly, system complexity and sophistication of inference did not correlate fully with performance, where some of the best results were obtained by rather naïve lexically-based systems. The fact that quite sophisticated inference levels were applied by some groups, with 6 systems applying logical inference, provides an additional indication that applied NLP research is progressing towards deeper semantic reasoning. Additional refinements are needed though to obtain sufficient robustness for the Challenge types of data. Further detailed analysis of systems performance, relative to different types of examples and entailment phenomena, are likely to yield future improvements.

Being the first benchmark of its types there are several lessons for future similar efforts. Most notably, further efforts can be made to create "natural" distributions of Text-Hypothesis examples. For example, *T*-*H* pairs may be collected directly from the data processed by actual systems, considering their inputs and candidate outputs. An additional possibility is to collect a set of multiple candidate texts that might entail a given single hypothesis, thus reflecting typical ranking scenarios. Data collection settings may also be focused on typical "core" semantic applications, such as QA, IE, IR and summarization. Some of these improvements are planned for the 2nd PASCAL Recognising Textual Entailment Challenge. Overall, we hope that future similar benchmarks will be carried out and will help shaping clearer frameworks, and corresponding research communities, for applied research on semantic inference.

Acknowledgements

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

- Document Understanding Conferences (DUC) 2004 Machine Translation evaluation data, from the National In-stitute of Standards and Technology (NIST). http://duc.nist.gov/duc2004/
- TextMap Question Answering online demo, from the Information Sciences Institute (ISI). http://brahms.isi.edu:8080/textmap/
- Relation Recognition dataset, from University of Illinois at Urbana-Champaign. http://l2r.cs.uiuc.edu/~cogcomp/
- DIRT paraphrase database (online demo), from the University of southern California. http://www.isi.edu/~pantel/Content/Demos/demosDirt.htm
- The output of the TEASE system for extracting entailment relations and paraphrases ([14]).
- Corpus of Sentence Alignment in monolingual comparable corpora, Columbia University. http://www.cs.columbia.edu/~noemie/alignment/

We would like to thank the people and organizations that made these sources available for the challenge.

We would also like to acknowledge the people involved in creating and annotating the data: Danilo Giampiccolo, Tracy Kelly, Einat Barnoy, Allesandro

Valin, Ruthie Mandel, and Melanie Joseph. 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, Florence d'Alché-Buc and Michele Sebag, for their enthusiastic efforts and assistance, and Eileen Simon and John Shawe-Taylor from PASCAL for being so supportive all along.

References

1. Bacchus, F.: Representing and Reasoning with Probabilistic Knowledge, M.I.T. Press (1990).
2. Bar-Haim, R., Szpektor, I., Glickman O.: Definition and Analysis of Intermediate Entailment Levels. ACL-05 Workshop on Empirical Modeling of Semantic Equivalence and Entailment. (2005)
3. Chierchia, G., McConnell-Ginet, S.: Meaning and grammar: An introduction to semantics, 2nd. edition. Cambridge, MA: MIT Press (2001).
4. Condoravdi, C., Crouch, D., de Paiva, V., Stolle, R., Bobrow, D.G.: Entailment, intensionality and text understanding. HLT-NAACL Workshop on Text Meaning (2003)
5. Corley,C., Mihalcea, R.: Measuring the Semantic Similarity of Texts. Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, pages 1318, Ann Arbor, June 2005.
6. Dagan, I., Glickman, O.: Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability. PASCAL workshop on Learning Methods for Text Understanding and Mining, 26 - 29 January (2004), Grenoble, France.
7. Glickman, O., Dagan, I., Koppel, M.: A Lexical Alignment Model for Probabilistic Textual Entailment. This Volums.
8. Halpern, J.Y.: An analysis of first-order logics of probability. Artificial Intelligence 46:311-350 (1990).
9. Keefe, R., Smith P. (ed.): Vagueness: A Reader. The MIT Press. 1997.
10. Lukasiewicz, J.: Selected Works, L. Borkowski Ed., North Holland, London, 1970.
11. Marsi, E., Krahmer, E.: Classification of Semantic Relations by Humans and Machines. Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, 2005.
12. Monz, C., de Rijke, M.: Light-Weight Entailment Checking for Computational Semantics. the third workshop on inference in computational semantics (ICoS-3). (2001)
13. Lucy Vanderwende and William B. Dolan: What Syntax can Contribute in the Entailment Task. This Volume.
14. Szpektor, I., Tanev, H., Dagan, I.,Coppola, B.: Scaling Web-based Acquisition of Entailment Relations. Empirical Methods in Natural Language Processing (EMNLP). (2004).
15. Zadeh, L.: Fuzzy sets. Information and Control, 8 , 1965.
16. Zaenen, A., Karttunen, L., Crouch, R.: Local Textual Inference: Can it be Defined or Circumscribed?. Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, 2005.