Natural Language Engineering **15** (4): i–xvii. © Cambridge University Press 2009
doi:10.1017/S1351324909990209

Recognizing textual entailment: Rational, evaluation and approaches

IDO DAGAN[1], BILL DOLAN[2], BERNARDO MAGNINI[3] and DAN ROTH[4]

[1]*Department of Computer Science, Bar Ilan University, Ramat Gan, 52900, Israel*
[2]*Natural Language Processing Group, Microsoft Research, One Microsoft Way, Redmond, WA 98005, USA*
[3]*Human Language Technologies Research Unit, Fondazione Bruno Kessler, Via Sommarive 18, 38050 Povo - Trento (Italy)*
[4]*Department of Computer Science, University of Illinois at Urbana-Champaign, IL, USA*
e-mail: dagan@cs.biu.ac.il, billdol@microsoft.com, magnini@fbk.eu, danr@uiuc.edu

(*Received 16 November 2007; accepted 6 February 2009*)

Abstract

The goal of identifying textual entailment – whether one piece of text can be plausibly inferred from another – has emerged in recent years as a generic core problem in natural language understanding. Work in this area has been largely driven by the PASCAL Recognizing Textual Entailment (RTE) challenges, which are a series of annual competitive meetings. The current work exhibits strong ties to some earlier lines of research, particularly automatic acquisition of paraphrases and lexical semantic relationships and unsupervised inference in applications such as question answering, information extraction and summarization. It has also opened the way to newer lines of research on more involved inference methods, on knowledge representations needed to support this natural language understanding challenge and on the use of learning methods in this context. RTE has fostered an active and growing community of researchers focused on the problem of applied entailment. This special issue of the *JNLE* provides an opportunity to showcase some of the most important work in this emerging area.

1 Introduction

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 (Dagan and Glickman 2004). This task 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 (MT) evaluation can be cast as entailment recognition (Dagan, Glickman and Magnini 2006). A necessary step in transforming

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press

textual entailment from a theoretical idea into an active empirical research field was the introduction of benchmarks and an evaluation forum for entailment systems. Dagan, Glickman and Magnini initiated in 2004 a series of contests under the PASCAL Network of Excellence, known as the PASCAL Recognising Textual Entailment (RTE) Challenges. These contests provided researchers concrete datasets on which they could evaluate their approaches, as well as a forum for presenting, discussing and comparing their results. The RTE datasets are freely available also for RTE non-participants, so as to further facilitate research on textual entailment.

This initiative has completed its fourth edition (Giampiccolo *et al.* 2008), now run under the U.S. National Institute of Standards and Technology (NIST), after three successful yearly PASCAL RTE Challenges – RTE1 in 2005 (Dagan, Glickman and Magnini 2005), RTE2 in 2006 (Bar-Haim *et al.* 2006) and RTE3 in 2007 (Giampiccolo *et al.* 2007).[1] In this frame, participating systems are required to judge the entailment value of short pairs of text snippets (termed the text T and the hypothesis H).

At the present time, RTE represents an important field of investigation. Beside the RTE Challenges, high interest in the natural language processing (NLP) research community is demonstrated by the organization of other workshops on the same topic, such as the Question Answering – Answer Validation Exercise (AVE) at Cross-Language Evaluation Forum (QA@CLEF 2008).[2] Concerning Italian, the second evaluation campaign of NLP tools for Italian (EVALITA 2009) has inserted textual entailment as one of its tasks,[3] following the terminology and the evaluation methods of the Pascal RTE Challenges for English.

This paper provides an overview of the relevant aspects in textual entailment. In Section 2 we provide the rationale behind textual entailment as a general approach to address language variability in applied semantics. Section 3 reviews the evaluation methodologies applied in the RTE campaigns. Section 4 is intended as a short overview of the approaches, the tools and the linguistic resources that have so far been used for textual entailment. Finally, Section 5 introduces the papers selected for this special issue.

2 Textual entailment: Rationale

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 NLP applications, such as QA, IE, (multi-document) summarization and 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.

[1] http://www.pascal-network.org/Challenges/{RTE,RTE2,RTE3,RTE4}
[2] http://nlp.uned.es/clef-qa/ave/
[3] http://evalita.fbk.eu/te.html

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press

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, 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 and communities have matured.

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 IR 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 expresses 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 automatic 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 entailment in formal semantics (Chierchia and Ginet 2001) 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 and 893 were annotated as 'True' even though the entailment in these cases is not absolutely certain. This seems to match the types of uncertain inferences that are typically expected from text-based applications. Glickman, Dagan and Koppel (2006) presented 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

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press

presupposition and implicature, see Zaenen, Karttunen and Crouch (2005). There is also a considerable amount of classical work on fuzzy or uncertain inference (e.g. Bacchus 1990; Halpern 1990; Keefe and Smith 1997). Making significant reference to this rich body of literature and deeply understanding the relationships between our definition of the operational textual entailment and the 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.

3 Textual entailment: Evaluation

The 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, the applied notion of *textual entailment* is defined as a directional relationship between pairs of text expressions, denoted by T the entailing 'text' and by H the entailed 'hypothesis'. We say that T *entails* H if, typically, a human reading T would infer that H is most probably 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 the evaluation of applied tasks such as QA and IE, 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 RTE1 dataset, along with their gold-standard annotation.

As in other evaluation tasks our definition of textual entailment is operational, and corresponds to the judgement criteria given to the annotators who decide whether this relationship holds for 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 Condoravdi *et al.* 2003; Dagan and Glickman 2004; Monz and de Rijke 2001).

The RTE systems' results demonstrate general improvement with time, with overall accuracy levels ranging from 50% to 60% on RTE1 (17 submissions), from 53% to 75% on RTE2 (23 submissions), from 49% to 80% on RTE3 (26 submissions) and from 45% to 74% on RTE4 (26 submissions, three-way task). Common approaches used by the submitted systems include machine learning (typically SVM), logical inference, cross-pair similarity measures between T and H and word alignment.

3.1 Collecting RTE datasets

The RTE datasets were designed by following (and supporting) the rationale that textual entailment recognition captures the underlying semantic inferences needed in many application settings. Accordingly, the text–hypothesis pairs were collected from several application scenarios, reflecting the way by which the corresponding application could utilize an automated entailment judgement. In other words, the

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press

Table 1. *Examples of text–hypothesis pairs, from the RTE1 dataset. The following abbreviations, corresponding to application settings, are used: QA for question answering, RC for reading comprehension, IR for information retrieval, MT for machine Ttranslation and CD for for comparable documents*

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
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 disappointing 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 fifth century AD.*	*The first settlements on the site of Jakarta were established as early as the fifth century AD.*	CD	True

derivation of entailment pairs within each application scenario mimics a potential reduction of the needed application inferences to the entailment recognition task. As a typical example of these processes, the following paragraphs describe the creation of entailment pairs under the four application scenarios in the RTE2 challenge.

3.1.1 Collecting information extraction pairs

This task is inspired by the IE (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 (that is to say the relations tested in the ACE-2004 RDR task) were taken as templates for hypotheses. Relevant news articles were collected as texts

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press

(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 fourth approach, hypotheses which correspond to new types of semantic relations (not found in the ACE and MUC datasets) were manually generated for sentences in the collected news articles. These relations were taken from various semantic domains, such as sports, entertainment and science. 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.

3.1.2 Collecting information retrieval 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, which served as the query.

3.1.3 Collecting question answering pairs

Annotators were given questions, taken from TREC-QA and QA@CLEF datasets, and the corresponding answers were extracted from the web by QA systems. Transforming a question–answer pair to a 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' 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. Similarly, a negative pair could have been generated by picking '5.5 thousand' as an (incorrect) answer term,

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press

resulting in the hypothesis 'Slovenia has 5.5 thousand inhabitants'. This process simulates the need of a QA system to verify that the retrieved passage text in which the answer was found indeed entails the provided answer.

3.1.4 Collecting summarization pairs

In this setting T and H are sentences taken from a news document cluster, which is 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 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 similarly simplified but this time without reaching the entailment of H by T. 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.

3.1.5 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 RTE1, pairs on which the annotators disagreed were filtered out. The average agreement on the test set (between each pair of annotators who shared at least 100 examples) was 89.2%, with an 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 because of disagreement.

Additional filtering was done by two of the organizers, who discarded pairs that seemed controversial, too difficult or redundant (or rather 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.

3.2 Evaluation measures

The main task in the RTE Challenges 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 optional task was *ranking* the pairs, according to their entailment confidence. In this ranking, the first pair is the one for which entailment is most certain, and the last pair is the one for which entailment is least likely (i.e. the one for which the judgement of 'no' is most certain). A perfect ranking would place all

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press

the positive pairs (for which the entailment truly 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 IR) 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 to say 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:

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

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 ith pair is positive[4] and 0 otherwise; and i ranges over the pairs, ordered by their ranking (note that this measure is different from the *confidence weighted score* used in RTE1).

3.3 RTE1

The first RTE challenge[5] (RTE1) introduced the first benchmark for textual entailment recognition (Dagan *et al.* 2005). The RTE1 dataset consisted of manually collected text fragment pairs, consisting of the *text* (T) (one or two 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 mentioned above. The dataset was split into a development set, containing 567 pairs, and a test set, containing 800 pairs. The pairs were balanced between positive (entailing) and negative (non-entailing) pairs.

The challenge raised noticeable attention in the research community, attracting 17 submissions from diverse groups. The relatively low accuracy achieved by the participating systems (best results below 60%) suggested that the entailment task is indeed a challenging one, with a wide room for improvement.

3.4 RTE2

Following the success and impact of RTE1, the main goal of the second challenge[6] (Bar-Haim *et al.* 2006) was to support the continuation of research on textual entailment. Four sites participated in the data collection and annotation: Bar-Ilan University (Israel, coordinator), CELCT (Italy), Microsoft Research (USA) and MITRE (USA). The main focus in creating the RTE2 dataset was to provide more 'realistic' text–hypothesis examples, based mostly on outputs of actual systems. As in the previous challenge, the main task was judging whether a hypothesis H is entailed by a text T. The examples represent different levels of entailment reasoning, such

[4] Note that in this formula, *positive* refers to the gold-standard annotation, not to the system's output.
[5] http://www.pascal-network.org/Challenges/RTE
[6] http://www.pascal-network.org/Challenges/RTE2

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press

as lexical, syntactic, morphological and logical. The data collection and annotation processes were improved, 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 some pre-processing for the examples was provided, including sentence splitting and dependency parsing.

3.5 RTE3

RTE3[7] (Giampiccolo *et al.* 2007) followed the same basic structure as RTE1 and RTE2, in order to facilitate the participation of newcomers and to allow 'veterans' to assess the improvements of their systems in a comparable setting. The main novelty of RTE3 was that part of the pairs contained longer texts (up to one paragraph), encouraging participants to move towards discourse-level inference. Twenty-six teams participated, and the results were presented at the ACL 2007 Workshop on Textual Entailment and Paraphrasing.

3.6 RTE4

The fourth challenge[8] (Giampiccolo *et al.* 2008) was co-organized, for the first time, by the NIST and was included as a track in the newly established Text Analysis Conference (TAC), together with QA and summarization. Hopefully, bringing these tracks together would promote the use of generic entailment technology within these and related applications.

The major innovation of the fourth challenge was a three-way classification of entailment pairs, piloted in RTE3. In three-way judgement, non-entailment cases are split between *contradiction*, where the negation of the hypothesis is entailed from the text, and *unknown*, where the truth of the hypothesis cannot be determined based on the text. Participants could submit to the three-way task, to the traditional two-way task or to both.

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

As regards the results, in the three-way task, the best accuracy was 0.685 calculated against the three-way judgement and 0.72 calculated against the 2-way judgement. The three-way task appeared to be altogether quite challenging, as the average three-way accuracy was 0.51, quite low compared to the results achieved in previous campaigns. The systems performed better in the two-way task, achieving accuracy

[7] http://www.pascal-network.org/Challenges/RTE3
[8] http://www.nist.gov/tac/tracks/2008/rte/

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press

scores which ranged between 0.459 and 0.746, with an average score of 0.573. These results are lower than those achieved in last year's competition, where the accuracy scores ranged from 0.49 to 0.80, even though a comparison is not really possible, as the datasets are different.

4 Textual entailment: Approaches, tools and resources

This section provides a short overview of the main approaches and resources used by textual Eentailment systems.

4.1 Approaches

The RTE task has fostered the experimentation of a number of data-driven approaches applied to semantics. Specifically, the availability of the RTE datasets for training made it possible to formulate the problem in terms of a classification task, where features are extracted from the training examples and then used by machine learning algorithms in order to build a classifier, which is finally applied to the test data to classify each pair as either positive or negative. The tendency to address textual entailment exploiting machine learning techniques is clearly showed in the RTE Challenges (see, among the others, Kozareva and Montoy 2006); Zanzotto, Pennacchiotti and Moschitti 2007), where most of the systems used machine learning algorithms (e.g. support vector machines) with a variety of features, including lexical-syntactic and semantic features, based on document co-occurrence counts, first-order syntactic rewrite rules, and to extract the information gain provided by lexical measures.

A related line of research addressing textual tntailment attempts to provide a number of transformations allowing to derive the hypothesis H from the text T. In this line, a number of transformation-based techniques over syntactic representations of T and H have been proposed, including the application of tree edit distance algorithms (Kouleykov and Magnini 2005) and the definition of transformation rules specifically designed to preserve the entailment relation for a number of language variability phenomena (Bar-Haim *et al.* 2008). The extension of transformation-based approaches towards a probabilistic setting is an interesting direction investigated by some systems.

The RTE task stimulated a number of approaches based on deep analysis and semantic inferences. In this direction both approaches based on logical inferences (see, among the others, Tau and Moldovan 2005; Bos and Markert 2006), the application of natural logic (Chambers *et al.* 2007) and approaches exploiting ontology-based reasoning, have been proposed and have often been coupled with data-driven techniques, where the final decision about the entailment relation is taken on the basis of semantic features managed by machine learning algorithms (de Salvo *et al.* 2005).

Finally, a recent direction is represented by precision-oriented RTE modules (see Wang and Neumann 2008; Cabrio, Kouylekov and Magnini 2008), where specialized textual entailment engines are designed to address a specific aspect

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press

of language variability (e.g. contradiction, lexical similarity), and then they are combined, applying a voting mechanism, with a high-coverage backup module.

4.2 Tools

Most of the approaches sketched in Section 4.1 do require intensive processing to be carried out over the T–H pairs, and the final results often crucially depend on the quality of the tools which are used. As a matter of fact, the ability to manage and combine good linguistic tools and resources has proved to be one of the key factors enabling high performance on the RTE datasets.

Mostly, tools have been used in order to pre-process data: tokenization, stemming, lemmatization and part-of-speech taggers are largely used, as well as parsers (e.g. Minipar, Stanford Parser) and named entity recognizers. In addition, software tools such as WEKA, for machine learning algorithms, and Lucene, for indexing and retrieval, have been largely used, as well as WordNet Similarity tools, aimed at estimating similarity measures on available lexical resources (see the next section for more details). More specific tools address anaphora resolution (relevant for long text, introduced in RTE3) and word sense disambiguation.

A relevant side effect of the RTE task is that several tools have been tested on the RTE datasets (see, among others, Iftene 2009), this way allowing evaluations and ablation tests, where the impact of a certain tool or resource has been qualitatively and quantitatively estimated.

4.3 Resources

Systems participating at the RTE evaluation campaigns have taken advantage of several available lexical-semantic resources. The most exploited of such resources is WordNet (Fellbaum 1998), with its extensions (e.g. EuroWordNet, eXtended WordNet). Also DIRT (Lin and Pantel 2001), a resource of statistically learned inference rules, has been used in several systems, as well as verb-oriented resources such as VerbNet (Kipper-Schuler 2005) and VerbOcean (Chklovski and Pantel 2004). The use of FrameNet (Baker *et al.* 1998) has been attempted by some systems, although in a limited way probably because of its restricted coverage or the difficulties in modelling FrameNet information. A renewed tendency in considering the web as a resource arises in the successful use of Wikipedia by some participating systems, in order to extract entailment rules, named entities and background knowledge.

Furthermore, various text collections are exploited as sources of information, such as the Reuters corpus and English Gigaword to extract features based on documents' co-occurence counts and InfoMap, Dekang Lins thesaurus and gazetteers to draw lexical similarity judgements.

A dedicated website, the Textual Entailment Resource Pool[9], hosted by ACL, provides a repository of linguistic tools and resources for textual entailment.

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

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press

4.4 Annotations and analyses of entailment phenomena

A relevant line of research, which is crucial for a deep understanding of textual entailment phenomena, is the development of annotated datasets, where *T–H* pairs are analysed according to the linguistic phenomena determining entailment. Several studies have tried to analyze such linguistic levels with respect to the entailment task, considering both lexical, syntactic and world knowledge.

As an example of *T–H* pairs' annotation, the ARTE dataset (Garoufi 2007) proposes a scheme for manual annotation with the aim of highlighting a wide variety of entailment phenomena in the data. ARTE views the entailment task in relation to three levels, i.e. *alignment*, *context* and *co-reference*, according to which 23 different features for positive entailment annotation are extracted. Each level is explored in depth for the positive entailment cases, while for the negative pairs a more basic and elementary scheme is conceived. The ARTE scheme has been applied to all positive entailment pairs in the RTE2 test set and to a random 25% sample of the negative pairs. As a result of the annotation, *reasoning* is the most frequent entailment feature appearing altogether in 65.75% of the annotated pairs: this indicates that a significant portion of the data involves deeper inferences. The combination of the entailment features is analyzed together with the entailment types and their distribution in the dataset.

An attempt to isolate the set of *T–H* pairs whose categorization as true or false can be accurately predicted based solely on syntactic cues has been carried out in Vanderwende and Dolan (2006). The aim of their work is to understand what proportion of the entailments in the RTE1 test set could be solved using a robust parser. Two human annotators evaluated each *T–H* pair of the test set, deciding whether the entailment was *true by syntax*, *false by syntax*, *not syntax* and *can't decide*. Additionally, annotators were allowed to indicate whether the information in a general purpose thesaurus entry would allow a pair to be judged true or false. Their results show that 37% of the test items can be handled by syntax, broadly defined; 49% of the test items can be handled by syntax plus a general purpose thesaurus. According to their annotators, it is easier to decide when syntax can be expected to return 'true', and it is uncertain when to assign 'false'. Basing on their own observations, the submitted system (Vanderwende, Menezes and Snow 2006) predicts entailment using syntactic features and a general purpose thesaurus, in addition to an overall alignment score. The syntactic heuristics used for recognizing false entailment heavily rely on the correct alignment of words and multi-word units between *T* and *H* logical forms.

Bar-Haim, Szpektor and Glickman (2005) defined two intermediate models of textual entailment, which correspond to lexical and lexical-syntactic levels of representation. Their lexical level captures knowledge about lexical-semantic and morphological relations and lexical world knowledge. The lexical-syntactic level additionally captures syntactic relationships and transformations, lexical-syntactic inference patterns (rules) and co-reference. They manually annotated a sample from the RTE1 dataset according to each model, compared the outcomes for the two models as well as for their individual components and explored how well they approximate

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press

the notion of entailment. It was shown that the lexical-syntactic model outperforms the lexical one, mainly because of a much lower rate of false-positives, but both models fail to achieve high recall. The analysis also showed that lexical-syntactic inference patterns stand out as a dominant contributor to the entailment task.

Referring to the RTE3 dataset, Clark and colleagues (Clark *et al.* 2007) suggested that only a few entailments can be recognized using simple syntactic matching and that the majority rely on significant amount of the so-called common human understanding of lexical and world knowledge. The authors presented an analysis of 25% of the RTE3 positive entailment pairs, to, one, identify where and what kinds of world knowledge are needed to identify and justify the entailment and, two, to discuss several existing resources and their capacity for supplying that knowledge (such as WordNet, the DIRT paraphrase database, FrameNet). After showing the frequency of the different entailment phenomena from the sample they analyzed, they stated that very few entailments depend purely on syntactic manipulation and simple lexical knowledge and that the vast majority of entailments require significant world knowledge.

A framework for semantic inference at the lexical-syntactic level is presented in Dagan *et al.* (2008), where an inference module is exploited for improving unsupervised acquisition of entailment rules through canonization (i.e. the transformation of lexical-syntactic template variations that occur in a text into their canonical form – this form is chosen to be the active verb form with direct modifier). The canonization rule collection is composed by two kinds of rules: (a) syntactic-based rules (e.g. passive/active forms, removal of conjunctions, removal of appositions) and (b) nominalization rules, trying to capture the relations between verbs and their nominalizations. The authors proposed to solve the learning problems using this entailment module at learning time as well.

Finally, a definition of contradiction for the textual entailment task is provided in de Marneffe, Rafferty and Manning (2008), together with a collection of contradiction corpora. Detecting contradiction appears to be a harder task than detecting entailment, since it requires deeper inferences, assessing event co-reference and model building. Contradiction is said to occur when two sentences are extremely unlikely to be true simultaneously (they must involve the same event). A previous work on the same topic was presented in Harabagiu, Hickl and Lacatusu (2006), in which the first empirical results for contradiction detection were provided (they focused only on specific kind of contradiction, i.e. those featuring negation and those formed by paraphrases).

5 Content of the special issue

This special issue of the *JNLE* presents five papers which well represent some of the main research areas in textual entailment. Zanzotto *et al.* and Harmeling present two foundational approaches to textual entailment, based, respectively, on machine learning and on probabilistic approaches. Zhao *et al.* address the problem of automatic acquisition of linguistic knowledge, specifically the acquisition of paraphrases from a bilingual corpus. Burchardt *et al.* investigate how textual entailment can

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press

benefit from already-developed semantic resources, specifically FrameNet. Finally, Nielsen *et al.* show how a textual entailment engine can be successfully used in a concrete application.

Zanzotto, Pennacchiotti and Moschitti ('A Machine Learning Approach to Textual Entailment Recognition') approach textual entailment under a machine learning perspective by first introducing the class of pair feature spaces, which allow supervised machine learning algorithms to derive first-order rewrite rules from annotated examples. In particular, they propose syntactic and shallow semantic feature spaces and compare them with standard ones. Extensive experiments demonstrate that the proposed spaces learn first-order derivations, while standard ones are not expressive enough to do so.

Harmeling ('Inferring Textual Entailment with a Probabilistically Sound Calculus') introduces a system for textual entailment that is based on a probabilistic model of entailment. The model is defined using a calculus of transformations on dependency trees, which is characterized by the fact that derivations in that calculus preserve the truth only with a certain probability. The calculus has been evaluated on the datasets of the PASCAL RTE Challenge.

Zhao, Wang, Liu and Li ('Extracting Paraphrase Patterns from Bilingual Parallel Corpora') present a pivot approach for extracting paraphrase patterns from bilingual parallel corpora, whereby the paraphrase patterns in English are extracted using the patterns in another language as pivots.

Burchardt, Pennacchiotti, Thater and Pinkal ('Assessing the Impact of Frame Semantics on Textual Entailment') underpin the intuition that frame-semantic information is a useful resource for modelling textual entailment. To this end, they provide a manual frame-semantic annotation for the test set used in the second RTE Challenge (i.e. the FATE corpus) and discuss experiments conducted on this basis. In particular, their experiments show that the frame-semantic lexicon provided by the Berkeley FrameNet project provides surprisingly good coverage for the task at hand. They identify issues of automatic semantic analysis components, as well as insufficient modelling of the information provided by frame-semantic analysis as reasons for ambivalent results of current systems based on frame semantics.

Finally, Nielsen, Ward and Martin ('Recognizing Entailment in Intelligent Tutoring Systems') present an application for textual entailment for intelligent tutoring. They describe a new method for recognizing whether a student's response to an automated tutor's question entails that they understand the concepts being taught. They demonstrate the need for a finer-grained analysis of answers than is supported by current tutoring systems or entailment databases and describe a new representation for reference answers that addresses these issues, breaking them into detailed facets and annotating their entailment relationships to the student's answer more precisely.

6 Conclusions

The RTE task has reached a noticeable level of maturity, as demonstrated by the very high interest in the NLP community and the gradually increasing participation in the RTE Challenges. Furthermore, the debates and the numerous publications

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press

about textual entailment have contributed to the better understanding of the task and its nature. We believe that this special issue is an important step in this direction.

As for the future of this area, we believe that the long-term goal of textual entailment research is the development of robust entailment 'engines'. Such engines will be used as a generic component in many text-understanding applications, encapsulating all the required semantic inferences, analogous to the use of part-of-speech taggers and syntactic parsers in today's NLP applications.

References

Bacchus, F. 1990. *Representing and Reasoning with Probabilistic Knowledge*. Cambridge, MA, USA, MIT Press.

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

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., and Szpektor, I. 2006. The second PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

Bar-Haim, R., Dagan, I., Mirkin, S., Shnarch, E., Szpektor, I., Berant, J., and Greenthal, I. 17 November 2008. Efficient semantic deduction and approximate matching over compact parse forests. In *Proceedings of the TAC 2008 Workshop on Textual Entailment*, Gaithersburg, Maryland, USA.

Bar-Haim, R., Szpektor, I., and Glickman, O. 2005. Definition and analysis of intermediate entailment levels. In *ACL-05 Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, University of Michigan, Ann Arbor, Michigan, USA.

Bos, J., and Markert, K. 2006 When logical inference helps determining textual entailment (and when it doesnt). In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

Cabrio, E., Kouylekov, M., and Magnini, B. 17 November 2008. Combining specialized entailment engines for RTE-4. In *Proceedings of the TAC 2008 Workshop on Textual Entailment*, Gaithersburg, MD.

Chambers, N., Cer, D., Grenager, T., Hall, D., Kiddon, C., MacCartney, B., de Marneffe, M.-C., Ramage, D., Yeh, E., and Christopher, D. M. 28–29 June 2007. Learning alignments and leveraging natural logic, In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic.

Chklovski, T., and Pantel, P. 2004. VerbOcean: Mining the web for fine-grained semantic verb relations. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, Barcelona, Spain.

Chierchia, G., and McConnell-Ginet, S. 2001. *Meaning and Grammar: An Introduction to Semantics*, 2nd ed. Cambridge, MA: MIT Press.

Clark, P., Harrison, P., Thompson, J., Murray, W., Hobbs, J., and Fellbaum, C. 28–29 June 2007. On the role of lexical and world knowledge in RTE3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic.

Condoravdi, C., Crouch, D., de Paiva, V., Stolle, R., and Bobrow, D. G. 2003. Entailment, intensionality and text understanding. *HLT-NAACL Workshop on Text Meaning*, Edmonton, Alberta, Canada.

Dagan, I., Bar-Haim, R., Szpektor, I., Greental, I., and Shnarch, E. 17–23 February 2008. Natural language as the basis for meaning representation and inference. In *Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing08)*, Haifa, Israel.

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press

Dagan, I., and Glickman, O. 2004. Probabilistic textual entailment: generic applied modeling of language variability, In *PASCAL Workshop on Learning Methods for Text Understanding and Mining*, Grenoble, France.

Dagan, I., Glickman, O., and Magnini, B. 11–13 April 2005.: The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment*, Southampton, UK.

Dagan, I., Glickman, O., and Magnini, B. 2006. The PASCAL Recognising Textual Entailment Challenge. In J. Quinonero-Candela, I. Dagan, B. Magnini, and F. d'Alch-Buc (eds.), *Evaluating Predictive Uncertainty, Visual Object Classification, and Recognizing Textual Entailment, First Pascal Machine Learning Challenges Workshop*, 2005, Berlin/Heidelberg, pp. 177–90, Lecture Notes in Computer Science, Vol. 3944, pp. 177–90. Springer-Verlag.

de Marneffe, M.-C. Rafferty, A. N., and Manning, C. D. 16–18 June 2008. Finding contradictions in text. In *Proceedings of the ACL 2008: HLT*, Columbus, OH.

de Salvo Braz, R., Girju, R., Punyakanok, V., Roth, D., and Sammons, M. 11–13 April 2005. An inference model for semantic entailment in natural language. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment*, Southampton, UK.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Garoufi, K. 2007. Towards a better understanding of applied textual entailment. *Master Thesis*. Saarbrücken, Germany: Saarland University.

Giampiccolo, D., Magnini, B., Dagan, I., and Dolan, B. 28–29 June 2007 The third PASCAL recognising textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic.

Giampiccolo, D., Trang, D, Hoa, Bernardo, M., Dagan, I., and Cabrio, E. 17 November 2008. The fourth PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the TAC 2008 Workshop on Textual Entailment*. Gaithersburg, MD.

Glickman, O., Dagan, I., and Koppel, M. 2006. A lexical alignment model for probabilistic textual entailment. In J. Quinonero-Candela, I. Dagan, B. Magnini, and F. d'Alch-Buc (eds.), *Evaluating Predictive Uncertainty, Visual Object Classification, and Recognizing Textual Entailment, First Pascal Machine Learning Challenges Workshop*, MLCW 2005, Lecture Notes in Computer Science Vol. 3944, pp. 287–98, Springer-Verlag.

Halpern, J. Y. 1990. An analysis of first-order logics of probability. *Artificial Intelligence* **46**: 311–50.

Harabagiu, S., Hickl, A., and Lacatusu, F. 16–20 July 2006. Negation, contrast, and contradiction in text processing. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, Boston, MA.

Iftene, A. 2009. *Textual Entailment*, PhD Thesis. Iasi, Romania: "Al. I. Cuza" University.

Keefe, R., and Smith, P. (ed.) 1997. *Vagueness: A Reader*. MIT Press.

Kipper Schuler, K. 2005. VerbNet: a broad-coverage, comprehensive verb lexicon. Dissertations available from ProQuest. University of Pennsylvania, Paper AAI3179808.

Kouleykov, M., and Magnini, B. 2005. Tree edit sistance for textual entailment. In *Proceedings of RALNP-2005, International Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria, pp. 271–8.

Kozareva, Z., and Montoyo, A. 2006. MLEnt: the machine learning entailment system of the University of Alicante. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

Landis, J. R., and Koch, G. G. 1997. The measurements of observer agreement for categorical data. *Biometrics* **33:** 159–74.

Lin, D., and Pantel, P. 2001 DIRT – Discovery of Inference Rules from Text. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-01)*, San Francisco, CA.

Monz, C., and de Rijke, M. 2001. Light-Weight entailment checking for computational semantics. In *The Third Workshop on Inference in Computational Semantics (ICoS-3)*, Siena, Italy.

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press

Szpektor, I., Tanev, H., Dagan, I., and Coppola, B. 2004. Scaling Web-based acquisition of entailment relations. In *2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, Barcelona, Spain.

Tatu, M., and Moldovan, D. 28–29 June 2007. COGEX at RTE 3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague: Czech Republic.

Vanderwende, L., and Dolan, W. B. 2006. What syntax can contribute in the entailment task. In J. Quinonero-Candela, I. Dagan, B. Magnini, and F. d'Alch-Buc (eds.), *Evaluating Predictive Uncertainty, Visual Object Classification, and Recognizing Textual Entailment, First Pascal Machine Learning Challenges Workshop*, MLCW 2005, Lecture Notes in Computer Science, Vol. 3944, pp. 205–16. Springer-Verlag.

Vanderwende, L., Menezes, A., and Snow, R. 10 April 2006. Microsoft research at RTE-2: syntactic contributions in the entailment task: an implementation. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

Voorhees, E. M., and Harman, D. 1999. Overview of the seventh text retrieval conference. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, Gaithersburg, MD.

Wang, R., and Neumann, G. 17 November 2008. An accuracy-oriented divide-and-conquer strategy. In *Proceedings of the TAC 2008 Workshop on Textual Entailment*, Gaithersburg, MD.

Zaenen, A., Karttunen, L., and Crouch, R. 2005. Local textual inference: can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, Ann Arbor, Michigan.

Zanzotto, F. M., Pennacchiotti, M., and Moschitti, A. 28–29 June 2007. Shallow semantic in fast textual entailment rule learners. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic.

https://doi.org/10.1017/S1351324909990209 Published online by Cambridge University Press