Chapter 26: Textual Entailment

Sebastian Padó and Ido Dagan

August 29, 2012

Abstract

Textual entailment is a binary relation between two natural language texts (the so-called text and hypothesis) that holds when readers of the text would agree that the hypothesis is most likely true (*Peter is snoring* $\rightarrow$ *A man sleeps*). The recognition of textual entailment requires an account of linguistic variability (i.e., the possibility to realize a certain state of affairs in different ways, as in *Peter buys the car* $\leftrightarrow$ *The car is purchased by Peter*) as well as of the derivation of additional knowledge (as in *Peter buys the car* $\rightarrow$ *Peter owns the car*). In contrast to classical (logics-based) inference, textual entailment also covers cases of very probable, but still defeasible, entailment (*A hurricane hit Peter's town* $\rightarrow$ *Peter's town was damaged*). A substantial part of human common-sense reasoning involves such defeasible inferences. As a consequence, textual entailment is of considerable interest for many real-world language processing tasks where it can serve as a generic, application-independent framework for semantic inference. This chapter discusses the history of textual entailment, relevant linguistic phenomena, approaches to recognizing textual entailment, and its integration in various NLP tasks.

26.1 Introduction: Inference and Entailment

Understanding the meaning of a text involves considerably more than reading its words and combining their meaning into the meaning of the complete sentence. The reason is that a considerable part of the meaning of a text is not expressed explicitly, but added to the text by readers through *(semantic) inference*:

> An *inference* is defined to be any assertion which the reader comes to believe to be true as a result of reading the text, but which was not previously believed by the reader, and was not stated explicitly in the text. Note that inferences need not follow logically or necessarily from the text; the reader can jump to conclusions that seem like but are not 100% certain. (Norvig 1987)

The drawing of such inferences, or "reading between the lines" (Kay 1987) is something that readers do instantaneously and effortlessly. It is a fundamental cognitive process that creates a representation of the text content that integrates additional knowledge the reader draws from his linguistic knowledge and world knowledge, and which establishes the coherence of the text. This enriched representation allows readers to, among other things, answer questions about texts and events that are not explicitly named in the text itself. For example, Norvig (1983) contrasts the following two examples:

(26.1) The cobbler sold a pair of boots to the alpinist.

(26.2) The florist sold a pair of boots to the ballerina.

From Example 26.1 on the preceding page, readers routinely draw the inferences that it was the cobbler who made the boots, and that the alpinist is buying the boots for the purpose of hiking in the mountains. These inferences are not present in Example 26.2, which is understood as a generic commercial transaction situation.

Mirroring Norvig's observations, Dagan and Glickman (2004) defined a notion of inference focused on text processing which corresponds to such **"common sense" reasoning patterns** under the name **Textual Entailment**. Textual Entailment is defined as a binary relation between two natural language texts (a **text** T and a **hypothesis** H) that holds if "a human reading T would infer that H is most likely true" where the truth of H could not be assessed without knowing T.

The ability to draw such inferences is relevant for a wide range of NLP text understanding applications such as Question Answering (QA), Information Extraction (IE), (multi-document) summarization, and the evaluation of Machine Translation (MT). From an application point of view, the drawing of inferences can also be characterized as a crucial strategy to deal with **variability in natural language**, that is, the fact, that the same state of affairs can typically be verbalized in many different ways. Variability can be considered as the dual problem of language ambiguity – the fact that many words in language have more than one meaning.

Figure 26.1 illustrates the role of inference on two examples. In the QA example (left-hand side), inference is used for answer validation. We assume that a QA system has some way to obtain a set of answer candidates. The role of inference here is to act as a filter: an answer candidate is considered

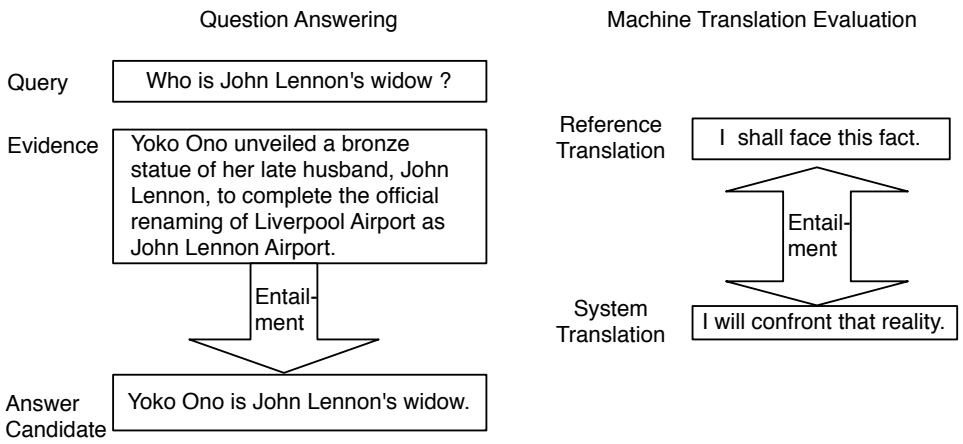

Figure 26.1: Inference in Text Understanding applications

correct only if it can be inferred from a sentence in the document. In the MT evaluation example (right-hand side), a human-provided reference translation and a system translation are given, and the question is whether the system matches the human translation well. This question can be mapped onto inference: A good system translation (as shown here) can be inferred from the reference translation, and vice versa. If the system translation contains spurious words, the inference from the reference to the system output fails, and if it is incomplete, it breaks down in the other direction.

An important observation about inferences in natural language is that they are generally defeasible: people not only draw inferences that are logically implied by their knowledge, but also inferences that are most likely true. In the case of Example 26.1 on page 2, the inference that the cobbler made the boots is plausible, but can be overridden. Still, in the absence of further information about the boots' provenance, we would expect an ideal QA system to return "the cobbler" for the question "Who made the boots

the alpinist bought?".

Textual Entailment can be contrasted with the classical *logical* concept of inference, such as the definition by Chierchia and McConnell-Ginet (2001), who state that T implies H if H is true in every possible world ("circumstance") in which T is also true. This definition of course implicitly relies on the possibility to formalize the meaning of T and H and then to determine the set of possible worlds in which either of them is true, problems that are far from being solved. In contrast, Textual Entailment is the call of a *human annotator* who assesses whether entailment holds. This decision naturally involves both linguistic knowledge and world knowledge. In consequence, Textual Entailment is neither a subset nor a superset of logical entailment. Logical entailments that are not Textual Entailments are, for example, all cases where the hypothesis is logically valid, i.e., a tautology (cf. the definition of Textual Entailment above). These cases are not considered Textual Entailments because the text does not contribute information towards the assessment of the hypothesis. Conversely, defeasible Textual Entailments, as discussed above, are not logical entailments.

The practical importance of Textual Entailment for Natural Language Processing lies in its potential to address the methodological problems of semantic inference methods for natural language. In this area, there is no clear framework of generic task definitions and evaluations. Semantic processing is often addressed in an application-oriented manner. It becomes difficult to compare practical inference methods that were developed within different applications, and researchers within one application area might not be aware of relevant methods that were developed for other applications. This

situation can be contrasted with the state of affairs in syntactic processing, where clear application-independent tasks have matured, and dependency structures are often used as a "common denominator" representation (Lin 1998b). The hope is that Textual Entailment can provide a similar service for the semantic processing field, by serving as a **generic semantic processing task** that forms a bridge between applications and processing methods. On the application side, the inference needs of many NLP tasks have been found to be reducible, at least to a large degree, to Textual Entailment. On the other side, Textual Entailment, being defined on the textual level, can be addressed with any semantic processing method. With regard to evaluation, Textual Entailment allows researchers to evaluate semantic processing methods in a representation-independent manner that, hopefully, is indicative of their performance in real-world natural language processing applications.

The remainder of this chapter is structured as follows. In Section 26.2, we introduce the Recognizing Textual Entailment (RTE) challenge, an annual shared task dedicated to textual entailment in the context of which much work has been done on developing the research area. Section 26.3 introduces linguistic phenomena that are relevant for textual entailment, and Section 26.4 gives an overview of computational strategies to modeling textual entailment and sources of knowledge about the relevant phenomena. Finally, Section 26.5 discusses the utility of textual entailment inference for various NLP tasks.

26.2 The Recognizing Textual Entailment Challenge

An important step in transforming Textual Entailment from a theoretical idea into an active empirical research field was the introduction of regular evaluation forums for Textual Entailment systems. In 2004, Dagan, Glickman and Magnini initiated a series of "shared task"-driven workshops under the PASCAL Network of Excellence, known as The PASCAL Recognising Textual Entailment Challenges (RTE). Since 2008, the RTE workshops have been organized under the auspices of the United States National Institute for Standards and Technologies (NIST), as one of the three tracks in the Text Analysis Conference (TAC). These contests have been providing researchers concrete datasets on which they could evaluate their approaches, as well as forums for presenting, discussing and comparing their results. The RTE datasets are freely available also for non-RTE participants in order to facilitate research on Textual Entailment.[1]

The RTE Challenges have changed in format over the years. RTE 1–3 (2005–7) focused on the fundamentals of the task, using individual sentences or very short, self-contained texts as texts and hypotheses and asking systems to make a binary entailment/non-entailment decision. Since this setup is a fairly limited proxy for NLP applications, subsequent years have extended the task in different respects. RTE 4 and 5 (2008/9) asked systems to make a *three-way* decision between entailment, non-entailments, and contradictions. The most recent events, RTE 6 and 7 (2010/11), adopt a new setup variously called "search" or "summarization" task. This section gives an overview of

ID	Text	Hypothesis	Task	TE?
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	yes
13	iTunes has seen strong sales in Europe.	Strong sales for iTunes software in Europe.	IR	yes
2016	Google files for its long awaited IPO.	Google goes public.	IR	yes
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	no

Table 26.1: Examples of Text-Hypothesis pairs from the RTE 1 dataset

data creation, the task extensions, and evaluation.

Early RTE Challenges. For each RTE workshop, new gold standard datasets were created by human annotation. For RTE 1 to RTE 3, the datasets consisted of both a development and a test set with 800 examples each. For RTE 4, only a test set was produced, with 1000 examples. RTE-5 resulted in development and test sets with 600 examples each. The data is organized according to "tasks", that is, subsets that correspond to typical success and failure cases in different typical applications (Dagan et al. 2009). In RTE 1, seven applications were considered; from RTE 2, this was reduced to four (Information Retrieval, Information Extraction, Question Answering, and Summarization). Table 26.1 shows examples from RTE 1.

The annotators were then presented with these text-hypothesis pairs, in

their original contexts, and asked to select an equal proportion (a 50%-50% split) of positive entailment examples, where T is judged to entail H, and negative examples, where entailment does not hold. They were asked to follow the somewhat informal definition of Textual Entailment from above. In this way, entailment decisions are based on (and assume) shared linguistic knowledge as well as commonly available background knowledge about the world.

Clearly, entailment decisions can be controversial. In particular, entailment may hinge on highly specific facts that are not commonly known, or annotators may disagree about the point at which a "highly plausible" inference that is accepted as a case of entailment becomes merely plausible. To gauge the size of this effect, all RTE 1 examples were tagged independently a second time. The first- and second-round annotation agreed for roughly 80% of the examples, which corresponds to a Kappa level of 0.6, or moderate agreement (Carletta 1996). 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% which might have been controversial. The final gold standard was very solid; partial re-annotations of participating teams showed agreement levels of between 91% and 96%. The annotation practices, including cross-annotation and reconciliation processes, were further refined along subsequent rounds of the RTE challenge.

Extension 1: Contradictions. In RTE 4 and 5, the task was extended by the introduction of *contradiction* as a third class. Contradictions are cases

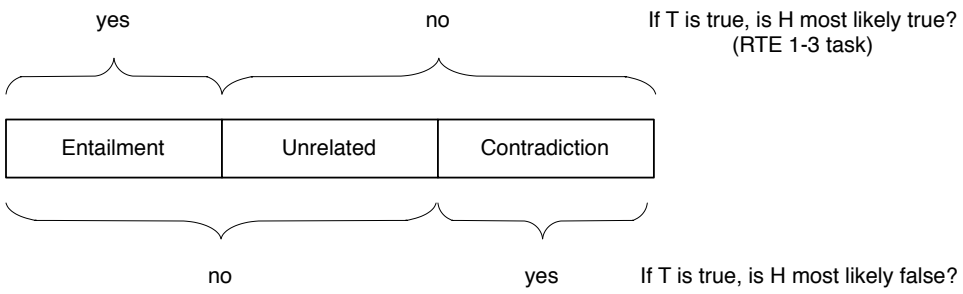

Figure 26.2: Three relations between sentence (or text) pairs: Entailment, Unrelated, Contradiction

where "assertions in H appear to directly refute or show portions of T to be false/wrong".[2] In parallel to the definition of entailment, contradictions do not need to be absolutely irreconcilable; a reconciliation "just has to appear highly unlikely in the absence of further evidence". The three-way split between entailment, contradiction, and unrelated cases is shown in Figure 26.2. The three-way decision can be seen as a refinement of the original binary decision, as shown by the curly brackets at the top, by splitting the "non-entailed" cases into unrelated and contradictory cases. (Note that the term "unrelated" here only refers to the textual entailment status, and not to topical relatedness.)

The introduction of contradiction also had ramifications for the design of the benchmark datasets. In RTE 1–3, the two classes (entailed/unrelated) were sampled to have an equal distribution, to obtain the lowest possible random baseline (50%). In practical applications, however, the "unrelated" class dominates the entailment class by far, while contradictions are much rarer than cases of entailment (De Marneffe et al. 2008). RTE 4 thus rejected the direct generalization to a three-class setup (equal numbers of entailments,

unrelated cases, and contradictions). Instead, the data was split into portions of 50% (entailed), 35% (unrelated) and 15% (contradiction).

Extension 2: Discourse context. RTE 4 also introduced longer texts that could consist of more than one sentence, forcing systems to integrate information from multiple sentences. Consider the following newswire article and corresponding query:

(26.3) *Text:* Chinese mining industry has always been plagued by disaster. [...] A recent accident has cost more than a dozen miners their lives.

Hypothesis: A mining accident in China has killed several miners.

While most of the hypothesis can be inferred from the second of these two sentences in the document, the crucial location information, as well as the fact that it is *mining* accidents which are discussed here, must be inferred from the first sentence, which requires at least some awareness of discourse structure.

Extension 3: The "Search" Task. In the most recent RTE challenges, RTE 6 and RTE 7, the task has been changed more fundamentally. Instead of determining semantic relations between text-hypothesis pairs in isolation, systems are presented with hypotheses and corresponding large sets of candidate texts, which are sentences embedded in complete documents. For each hypothesis, the systems must identify *all* sentences from among the candidate texts that entail the hypothesis. Typically, this decision has to

take context into account, like in Example 26.3 on the preceding page (Mirkin et al. 2010). This setup, which was first tested at RTE 5 as a pilot task, was explicitly designed to better model the application of Textual Entailment in NLP tasks (Bentivogli et al. 2009a). For example, it can be seen as a proxy for search tasks when the candidates are extracted from a large corpus with Information Retrieval methods. When the candidates are selected from among machine-created summaries of the text, the setup mirrors the validation of summarization output.

The new "search task" setup requires new reference datasets. Both RTE 6 and RTE 7 created data sets covering 20 topics (10 for development, 10 for testing) on the basis of existing Text Analysis Conference (TAC) Summarization corpora. For each topic, up to 30 hypotheses were selected, and for each hypothesis up to 100 candidate texts were annotated as entailing or non-entailing. This resulted in development sets with 16,000 (RTE 6) and 21,000 (RTE 7) and test sets with 20,000 (RTE 6) and 22,000 (RTE 7) sentences, respectively. See Bentivogli et al. (2010) and Bentivogli et al. (2011) for details.

Evaluation. In the first RTE challenges, systems simply returned the set of examples for which entailment held. This output was evaluated using accuracy (ratio of correctly classified examples). Later, systems were given the ability to *rank* examples according to their confidence in the entailment. The Average Precision (AP) statistic (Salton and McGill 1983) evaluates this type of output by giving higher-ranked examples a higher weight in the result. AP is defined as an average over the precision values for all top-n ranked lists

which have a true positive in its last position. For the RTE 4/5 three-class setup, Bergmair (2009) also proposed an information-theoretic evaluation metric that takes class imbalances into account (cf. above). Finally, the search setup of RTE 6/7 has been evaluated with classical Information Retrieval evaluation measures, namely Precision, Recall, and F_1 score.

26.3 Entailment Phenomena

In the collection of the RTE datasets, no effort was made to select or even mark examples by the types of linguistic phenomena or types of knowledge that they include. The rationale was to stay as close as possible to the needs of practical NLP applications, which typically encounter examples where any number of phenomena and types of knowledge interact. Thus, even perfect mastery of some phenomena may not make a large difference in the performance on the RTE dataset, and it is hard to translate the performance of a system into assessments about its grasp of semantic phenomena. This observation has sparked an early debate (Zaenen et al. 2005; Manning 2006; Zaenen et al. 2006). Since then, a number of studies have analyzed the Textual Entailment datasets with regard to the phenomena that they involve and have proposed different classifications (Vanderwende and Dolan 2006; Bar-Haim et al. 2005; Clark et al. 2007; Garoufi 2008).

In this section, we give a rough overview of important linguistic phenomena in Textual Entailment that is intended to indicate the challenges involved in this task. Following the intuition from Section 26.1 that entailment has to address variability in verbalization, Table 26.2 shows a simple classification

	Syntax/Morphology	Semantics
Words	Derivations, Abbreviations	Ontological and world knowledge-based lemma-level relations, discourse reference
Phrases	Alternative syntactic constructions	Ontological and world knowledge-based phrase-level relations, argument structure alternations

Table 26.2: Types of Entailment Phenomena: Classification by the level on which transformation takes place and by the type of knowledge involved

of the different types of linguistic differences (or transformations) we see between texts and hypotheses. They are separated into rows based on whether they pertain to individual words or to larger phrases, and into columns based on linguistic levels.

Note that this enumeration of entailment phenomena does not imply that systems without world knowledge are in principle unable to correctly recognize entailments. Rather, these phenomena can often be approximated with shallow cues (morphological, syntactic, or lexical), as it is the case in many other tasks in natural language processing. To the extent that this is possible, entailment can be decided with only limited linguistic knowledge; cf. Section 26.4 for details.

Derivations, Abbreviations. This class consists of transformations that account for differences between expressions that can be used alternatively to refer to the same entities or events. These include morphological transformations such as nominalizations (*sell* vs. *sale*) and abbreviations (*Mr Bush* vs. *George Bush*, *USA* vs. *United States*).

Alternative syntactic constructions. This class is composed of transformations that reflect general choices in the surface realization while keeping the lexical elements and the semantic relationships between them constant. For example, coordination, appositions, and relative clauses can often be used interchangeably: *Peter, who sleeps soundly, snores* means the same thing as *Peter, sleeping soundly, snores* or *Peter sleeps soundly and snores*. Another prominent instance is formed by English genitives which can often be expressed either as *X of Y* or *Y's X*.

Lemma-level relations based on ontological and world knowledge. This class concerns lemma-specific transformations at the level of individual words. As ontological knowledge, we consider transformations that instantiate a clear lexical relation such as synonymy, hypernymy/hyponymy, or meronymy/holonymy for nouns (*table* → *furniture*) or troponymy ("X is a manner of Y") or for verbs. World knowledge covers everything above and beyond ontological knowledge, such as knowledge about causation relations (*kill* → *die*), temporal inclusion (*snore* → *sleep*), or knowledge about named entities (in the case of ID 13 in Table 26.1, the fact that iTunes is a software product). Clark et al. (2007) give a detailed analysis of this category, distinguishing, for example "Core Theories" (about spatial reasoning or set membership) from "Frame/Script Knowledge".

Discourse references. Determining inference often involves the resolution of discourse references such as coreference or bridging, as the example above (*her* → *Yoko Ono's*) has shown (Mirkin et al. 2010).

Argument structure alternations. In this class, a predicate remains the same, but the realization of its arguments changes. The most prominent phenomenon in this class is passivization, which involves the promotion of the original object to a subject, and the deletion (or demotion to a *by*-PP) of the former subject. Many predicates can also show other types of diathesis alternations, such as the double object alternation (*Peter gave the book to Mary* → *Peter gave Mary the book*).

Phrase-level relations based on ontological and world knowledge. This class is the phrase-level analogue to ontological and world knowledge-based relations at the word level. It comprises transformations that combine a modification of the syntactic structure with modifications of their lexical elements. Such phrasal relations are ubiquitous in the RTE datasets and are maybe the class that clearest reflects the natural ability of language users to verbalize states of affairs in different ways, or to draw inferences from them (cf. Section 26.1). Phrasal relations can be asymmetrical or symmetrical. The term *paraphrase* is often used to refer to both types (cf. Section 26.4.2), even though strictly speaking it is only appropriate for the symmetrical case.

In the QA example in Figure 26.1, we have used the paraphrase (symmetrical phrase-level relation) *X is late husband of Y* ↔ *Y is widow of X*. Another symmetrical example is *X files for IPO* ↔ *X goes public* (ID 2016 in Table 26.1). An example of a non-symmetrical phrasal relation, where entailment only holds in one direction, would be the event/post condition inference *X kill Y* → *Y die*.

Monotonicity. Up to now, the entailment transformations in this section have been presented as if they could be applied to words or phrases regardless of context. However, the applicability of many transformations is subject to the *monotonicity* (upward or downward) of the context (Nairn et al. 2006; MacCartney and Manning 2007, 2008). A prominent example is deletion. In upward monotone contexts – such as main clauses without negation – material can be freely deleted. In contrast, this is not true in downward monotone contexts, which can for example be introduced by particles (*no/not*), some embedding verbs, some quantifiers like *few*, or even just by superlative constructions. Contrast:

(26.4) *Fido is a black terrier.* $\Rightarrow$ *Fido is a terrier.* (upward monotone)

(26.5) *Fido is not a black terrier.* $\not\Rightarrow$ *Fido is not a terrier.* (downward monotone)

(26.6) *Fido is the smallest black terrier.* $\not\Rightarrow$ *Fido is the smallest terrier.* (downward monotone)

Monotonicity equally influences the relationship between ontological relations and entailment: The replacement of words by their hypernyms works only in upward monotone contexts (*table* $\rightarrow$ *furniture*), but not in downward monotone ones, where words must be replaced by their hyponyms (*furniture* $\rightarrow$ *table*).

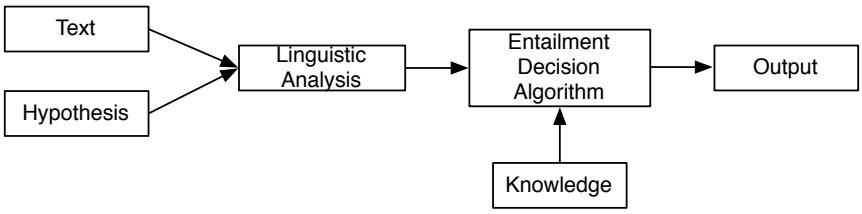

Figure 26.3: Structure of an entailment engine

26.4 Building Entailment Engines

We now come to the question of how the entailment decision is computed by practical entailment engines. Figure 26.3 shows the typical overall structure of such engines. Virtually all current systems perform some linguistic analysis of the input. This can include segmentation, stemming, lemmatization and part-of-speech tagging, as well as parsing and named entity recognition (see Chapters 21 to 25, 27 and 28 for details on processing methods). The ability to manage and combine good linguistic tools and resources has shown to be a key factor for high performance in the RTE challenges, since the normalization of the linguistic input already abstracts away from a certain amount of spurious surface variability.

After analysis, text and hypothesis are handed over to the entailment decision algorithm which determines whether entailment holds, typically taking into account various knowledge sources. We distinguish three main groups of entailment decision algorithms, namely (a), matching-based algorithms; (b), transformation-based algorithms; and (c), logics- or knowledge representation-based algorithms. The difference between (a) and (b) on one side and (c) on the other side is that (a) and (b) operate directly on linguistic

representations, while (c) translates the input into a formal language. (a) is distinguished from (b) in that (a) directly compares T and H, while (b) attempts to transform T into H, possibly through a number of intermediate steps. Clearly, these three classes do not exhaust the space of possible entailment decision algorithms, but they provide a convenient classification of existing approaches.

There are two further parameters which are orthogonal to the class of the entailment decision algorithms. The first one is the concrete processing framework. Each class of algorithms has the ability to use, for example, machine learning, rule-based, or heuristic frameworks. In practice, most systems use supervised machine learning due to its robustness and ability to deal with noisy and uncertain information. The second parameter is the knowledge used in the decision. Again, algorithms from all classes can employ the same kind of lexical, syntactic, or world knowledge (cf. Section 26.4.2).

26.4.1 Entailment Decision Algorithms

Matching-based Entailment. The first major class of entailment decision algorithms establishes a *match* or *alignment* of some sort between linguistic entities in the text and the hypothesis. It then estimates the quality of this match, following the intuition that in cases of entailment, the entities of H can be aligned well to corresponding linguistic entities in T, while this is generally more difficult if no entailment holds.

In one of the earliest approaches to Textual Entailment, Monz and de Rijke (2001) applied this idea to a bag of words (BoW) representation by simply computing the intersection of the BoW for text and hypothesis and

comparing it to the BoW for the hypothesis only. This scheme turned out to be effective, obtaining an accuracy of 58% on RTE 1, comparable to the best result of any system participating on the RTE 1 challenge. Nevertheless, the correlation between entailment and word overlap is not perfect – this method cannot deal either with text-hypothesis pairs that differ in few, but crucial, words (false positives), or which conversely involve much reformulation (false negatives). Much subsequent work has therefore experimented with matching and alignment on actual linguistic structure of the input, such as dependency trees, frame-semantic graphs (Burchardt et al. 2009), words (Hickl and Bensley 2007) or non-hierarchical phrases (MacCartney et al. 2008). Alignments are generally established using various kinds of knowledge (lexical, syntactic, phrasal) from knowledge sources like WordNet or thesauri (cf. Section 26.4.2). Zanzotto et al. (2009) employ tree kernels to encode first-order rewrite rules on constituency trees.

The easiest way to decide entailment is to directly use statistics of the alignment, such as the lexical coverage of H or its sum of edge weights, as features in some classifier that decides entailment. Burchardt et al. (2009) find that the frame-semantic structures of H in fact align substantially better with T when T entails H than if it does not. However, MacCartney et al. (2006) present a series of arguments against using alignment quality as the sole proxy of entailment probability. First, simple matching approaches generally ignore unaligned material. This corresponds to an assumption of upward monotonicity that is not generally valid (consider, for example, unaligned negations). Second, alignment computation must typically be broken down into local decisions (taking limited context into account) to be

feasible. This method can only do limited justice to non-local phenomena like polarity or modality. Third, the matching approach attempts to recognize cases of non-entailment through low-scoring alignments. At the same time, the alignments generally result from a search whose goal is to identify a high-scoring alignment. Thus, systems tend to avoid low-scoring alignments wherever possible and identify instead "loose" correspondences between material in T and H (see MacCartney et al. for examples).

To avoid these problems, McCartney et al. propose a two-stage architecture. In the first stage, an optimal alignment is computed from local alignment scores for words and edges. The second stage constructs a set of features that represent "entailment triggers", i.e., small linguistic theories about properties of entailing and non-entailing sentences. Their implementation, the Stanford Entailment Recognizer (MacCartney et al. 2006), takes a range of features into account, including factivity, polarity, modality, matches of names and numbers, syntactic compatibility, and alignment quality. The features form the basis of a linear classifier that decides entailment.

Other extensions to the basic matching paradigm include Hickl and Bensley (2007), who extract "discourse commitments" (atomic propositions) from T and H and check whether each H commitment is entailed by a T commitment with simple lexical scoring, assuming that the simpler structure of commitments alleviates the limitations of matching. Shnarch et al. (2011) addresses the limitations of matching in a different manner, by learning a probability model for lexical entailment rules from different resources that makes it possible to compute well-defined entailment probabilities for H-T pairs.

Transformation-based Entailment. A second class of approaches operationalized entailment by concentration on the existence of a "proof", i.e., a sequence of meaning-preserving transformation steps that converts the text into the hypothesis (cf. Section 26.3). Formally, a proof is a sequence of consequents, $(T, T_1, \ldots, T_n)$, such that there is an n with $T_n{=}H$ (Bar-Haim et al. 2009; Harmeling 2009), and that in each transformation step, $T_i \rightarrow T_{i+1}$, the consequent T_{i+1} is entailed by T_i. The main contrast to matching is that transformation-based approaches are able to model sequences of transformations (*chaining*).

A particularly simple family of transformation approaches is based on tree edit distance. These approaches generally define sets of globally applicable tree edit operations. The costs of these operations can be estimated either by parametrizing them based on linguistic properties of the tree (Harmeling 2009; Wang and Manning 2010) or without drawing on linguistic knowledge (Negri et al. 2009; Heilman and Smith 2010).

Many transformation-based systems take a more knowledge-intensive approach. They require that steps in the proofs they construct are validated by *inference rules*. Like in the case of matching, these rules can describe entailments on different levels and can be drawn from various knowledge sources.

An example of such a system is the first version of the Bar-Ilan University Textual Entailment Engine BIUTEE (Bar-Haim et al. 2007, 2009). In a first step, BIUTEE applies general inference rules for syntax and argument structure, such as passive-active transformations and the extraction of embedded propositions or appositions (cf. Table 26.2). This step relies on a relatively

small set of hand-written syntactic inference rules whose applicability is checked with a simple polarity model (Bar-Haim et al. 2007). The second step applies semantic inference rules, both at the level of individual words and at the phrasal level (cf. Table 26.2), to integrate ontological and world knowledge. Due to the lexically specific nature of the information used in this step, the resources for this step are very large, containing hundreds of thousands to millions of rules (cf. Section 26.4.2). This leads to serious efficiency problems, since most inference rules are independent, and thus the number of derivable consequents grows exponentially in the number of rule applications. To alleviate this problem, BIUTEE does not search the space of all possible rewrites of T naively, but computes a *compact parse forest* representation of the consequents of T. Figure 26.4 shows an example. The short text on the left-hand side, combined with three inference rules, leads in $2^3 = 8$ possible hypotheses (including the text itself). The compact parse forest, which subsumes all possible consequents, is shown on the right-hand side.

A practical problem of this type of system is caused by the imperfect coverage of the knowledge resources. Ideally, the system would predict entailment if and only if there is a proof from T to H. In practice, however, it is often impossible to find a complete proof even if entailment holds. As an example, the search-based transformation system of Dinu and Wang (2009), which represents T and H with dependency trees, constructs proofs from paraphrase rules, and uses this decision procedure, produces a high precision on examples within its coverage, but can only cover about 10% of all examples.

Text: *Children are fond of candies*

Inference rule 1: *candies* → *sweets*
Inference rule 2:*children* → *kids*
Inference rule 3: *X is fond of Y* → *X likes Y*

Hypothesis 1: *Kids are fond of candies*
Hypothesis 2: *Kids are fond of sweets*
Hypothesis 3: *Children like sweets*
Hypothesis 4: *Kids like sweets*
...

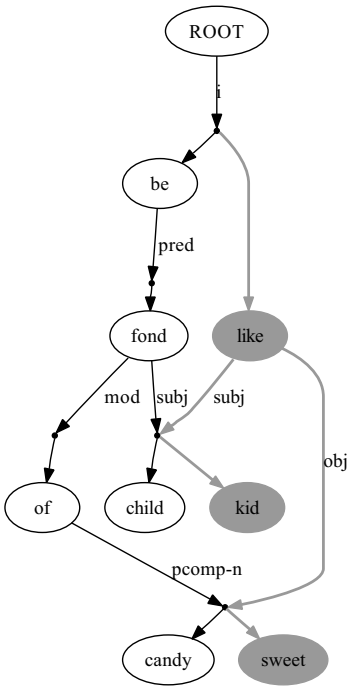

Figure 26.4: Left hand side: Short text with three inference rules and 4 of 8 derivable consequents. Right hand side: Compact parse forest representation of the consequents.

To close this gap, many transformation-based systems add a final step that compares the text's consequents and the hypothesis using classification or similarity-based matching techniques as described abve (Bar-Haim et al. 2009; de Salvo Braz et al. 2005). This approach yields a hybrid system that includes both a knowledge-based transformation component and a similarity-based matching component.

Taking an alternative approach, the second version of BIUTEE (Stern and Dagan 2011) extends the transformation-based paradigm to combine knowledge-based transformations with tree edits to alleviate the need for

matching. Their system applies a coherent transformation mechanism that always generates the hypothesis from the text. It does so by adding ad-hoc tree-edit transformations of pre-defined types, such as substitution, insertion, or relocation of nodes or edges. While these ad-hoc transformations are not grounded in pre-defined knowledge, the likelihood of their validity is heuristically estimated based on various linguistically-motivated features. An iterative learning algorithm estimates the costs of all types of transformations, both ad-hoc and knowledge-based, quantifying the likelihood that each transformation preserves entailment. Then, the optimal (lowest-cost) mixture of operations, of all types, is found via an Artificial Intelligence-style heuristic search algorithm (Stern et al. 2012). This second version of BIUTEE is available as open source (Stern and Dagan 2012).[3]

Logics/Knowledge Representation-based Entailment The final class comprises approaches which represent text and hypothesis in some formal language. These range from less expressive description logics (Bobrow et al. 2007) to first-order logics (Bos and Markert 2005; Raina et al. 2005; Tatu and Moldovan 2005). The use of formal languages is motivated by the availability of provably correct inference mechanisms. Consequently, text-hypothesis pairs for which H can be proven given T are virtually certain entailments, while H and T are contradictory if their formulae are not simultaneously satisfiable. Unfortunately, strict realizations of this approach tend to suffer from low recall, since for most H-T pairs the two formulae are neither entailing nor contradictory. The reason is that the broad-coverage construction of logics-based semantic representations for natural language sentences

must deal with tricky issues such as multi-word expressions, intensionality, modalities, quantification, etc., and is essentially an open problem. Also, the background knowledge must be represented in the same language, typically in the form of meaning postulates. For this reason, most approaches introduce some form of generalization or relaxation. Raina et al. (2005) automatically learn additional axioms necessary to complete proofs. Bos and Markert (2005) compute features over the output of a logics-based inference system and train a classifier on these features.

MacCartney and Manning (2008) adopt a weak formal calculus, *natural logic*, a system of logical inference which operates directly on natural language. Natural logic avoids many of the problems involved in traditional syntax-semantics interfaces but still supports a considerably more precise assessment of entailment than proposed so far within transformation- or alignment-based systems, in particular with regard to polarity. However, MacCartney and Manning's calculus cannot deal well with other aspects of inference frequent in the RTE data, like temporal reasoning or multi-word paraphrasing.

26.4.2 Knowledge Sources

The three classes of entailment decision algorithms sketched in Section 26.4.1 use knowledge about entailment phenomena in different ways. Nevertheless, on the representational side, all of them can be seen as using **inference rules** that describe possible local inference steps, possibly with conditions attached that determine their applicability, or with scores which describe their quality or reliability. Analyses performed on the output of RTE systems, as well as dedicated feature ablation tests, have consistently shown the crucial

importance of high-quality knowledge resources, as well as the challenge of representing the knowledge in the shape of effective features.

Many of the resources commonly used in Textual Entailment systems are general-purpose resources that have been applied to many other NLP tasks. A standard choice that is included in almost all systems is WordNet (Fellbaum 1998), which provides semantic information on the word level (cf. Table 26.2) in the form of a hand-constructed deep synonymy- and hyponymy-based hierarchy for nouns that is extended with other relations, and a flatter verb hierarchy constructed around different types of entailment. WordNet has been extended in various directions, such as increased coverage (Snow et al. 2006), formalization of synset meaning (Moldovan and Rus 2001; Clark et al. 2008), and addition of argument mappings for verbs (Szpektor and Dagan 2009). See Chapter 20 for more details.

Other resources provide deeper information, such as verb classes and semantic roles. Prominent examples are VerbNet (Kipper et al. 2000) and FrameNet (Fillmore et al. 2003). Such resources have also been used in entailment, although it has been found that knowledge from semantic roles, which is more concerned with "aboutness" than with truth values, needs to be enriched with other types of evidence (Ben Aharon et al. 2010)

The major shortcoming of such hand-constructed resources is their limited coverage. Therefore, the extraction of semantic knowledge from large corpora with machine learning methods has received a great deal of attention (see also Chapters 12 and 13). The simplest type of knowledge is symmetric semantic similarity computed with distributional methods, for which Lin's thesaurus (Lin 1998a) is an example. Kotlerman et al. (2010) present an

asymmetrical semantic similarity measure the output of which can be interpreted as inference rules. Chklovski and Pantel (2004) use surface patterns to identify verb pairs from different semantic relations, Danescu-Niculescu-Mizil and Lee (2010) learn downward monotone operators, and Shnarch et al. (2009) acquire inference rules form Wikipedia. At the phrasal level, most approaches induce paraphrase relations, either among strings (Bannard and Callison-Burch 2005) or among syntax tree fragments (Lin and Pantel 2002; Szpektor et al. 2004; Zhao et al. 2008; Szpektor and Dagan 2008). An exception is Zanzotto et al. (2009) who induce proper asymmetrical inference rules at the phrasal level from entailing sentence pairs.

Mirkin et al. (2009a) analyze the contribution of individual knowledge sources and find that all state-of-the-art resources are still lacking with respect to both precision and recall. With regard to precision, it has been proposed to verify the applicability of rules in the knowledge sources for each new instance based on the context of the proposed local inference (Pantel et al. 2007; Szpektor et al. 2008).

Last, but not least, similar machine learning methods have been applied to the related but separate task of acquiring additional entailing text-hypothesis pairs from large corpora. Hickl et al. (2006) extract 200,000 examples of entailing text-hypothesis pairs from the WWW by pairing the headlines of news articles with the first sentence of the respective article. Bos et al. (2009) construct a new Textual Entailment dataset for Italian guided by semi-structured information from Wikipedia.

26.5 Applications

As discussed in Section 26.1, the potential of Textual Entailment lies in providing a uniform platform which can be used by a range of semantic processing tasks in a manner similar to the use of a generic parser for syntactic analysis. The pivotal question is of course the **reduction to entailment** for each application: what part of the task can be solved by entailment, and how can it be phrased as an entailment problem?

26.5.1 Entailment for Validation

The first large class of applications uses entailment as validation. The tasks in this class are typically retrieval tasks, such as Information Extraction (cf. Chapter 35) or Question Answering (cf. Chapter 36). For these tasks, some query is given, and text that is relevant for the query is to be identified in a data source. In an ideal world the retrieval tasks could be mapped completely onto entailment: the query corresponds to the hypothesis, and the sentences of the data source correspond to texts (cf. Figure 26.1). Each text that entails the hypothesis is returned. Unfortunately, the data sources are often huge (such as the complete WWW), and it is typically infeasible to test all sentences for entailment with the query. This suggests a two step procedure. The first step is *candidate creation*, where a set of possible answers is computed, mostly with shallow methods. The second step is *candidate validation*, where the most suitable candidates are determined by testing entailment.

Question Answering. In Question Answering, the query consists of a natural language sentence like "Who is John Lennon's widow?". The candidate creation step consists of document and passage retrieval, which are typically based on IR methods that represent documents in vector space. Entailment is used in the second step, candidate validation. The most widely used strategy is to test for entailment between the retrieved passage (as T) and a hypothesis H that is obtained by turning the question into a declarative sentence, replacing the question word with a gap or variable. Peñas et al. (2008) employed this idea to multiple languages in the context of the CLEF conference. Observing that the best individual system only returns a correct answer for 42% of the questions, although over 70% of the questions could be answered by at least one participant, they emphasize the potential of Textual Entailment to abstract away from individual processing paradigms. However, the conversion of questions into statements can also lead to wrong results: For the query *Where did the Titanic sink?*, the hypothesis *The Titanic sank in* __ is entailed by *The Titanic sank in the North Atlantic* (correct answer), but also by *The Titanic sank in 1912* (incorrect answer).

Harabagiu and Hickl (2006) avoid this problem by reversing the process: instead of turning the question into a declarative sentence, they automatically generate, for each candidate passage, the set of all questions that can be answered from it. They then test for Textual Entailments between these questions and the original question.

Celikyilmaz et al. (2009) address the problem of data sparseness posed by the small size of the labelled entailment datasets produced by RTE and propose semi-supervised graph-based learning to acquire more example pairs

of likely and unlikely entailment.

Relation Extraction. In relation extraction, the queries are templates with slots like *A approaches B*. The goal is to identify sentences that instantiate these templates in text. This time, the query can be interpreted directly as a hypothesis, and the goal is to identify texts that entail the hypothesis. Roth et al. (2009) argue that the best strategy for high-recall high-precision relation extraction is to phrase the problem as an entailment task. Romano et al. (2006) investigate the adaptation of a relation extraction system to the biomedical domain, and find the largest problem to be the identification of domain-appropriate paraphrasing rules for the templates (cf. Section 26.3). Since the corpus was small enough, they did not require a "candidate creation" step, but were able to directly apply all known inference rules to each sentence to match it against the relations of interest. Bar-Haim et al. (2007), who target a large corpus, apply inference rules backward to relation templates in order to create shallow search engine queries for candidate creation. The returned snippets were parsed, and those for which the validity of the proof could be confirmed were accepted. In an error analysis, it was found that the inclusion of lexicalized semantic inference rules increased the recall six-fold compared to just syntactic inference rules, but precision dropped from 75% to 24%, indicating again that inference rules must be tested for applicability in context.

26.5.2 Entailment for Scoring

Scoring tasks form a second class. Here, usually two sentences or paragraphs of interest are provided, a candidate and a gold standard. The desired outcome is a judgment of their semantic relationship. This answer can be binary (entailed/non-entailed), but often a graded assessment of the *degree of semantic equivalence* between the candidate and the gold standard is desired.

Intelligent Tutoring. The goal of scoring in intelligent tutoring is to assess the quality of a student answer, given a gold standard answer (cf. Chapter 43 for details). Previously, this task was usually approached by building a so-called *model* of the correct answer, that is, a large number of possible realizations, so that student answers could be matched against the model. Textual Entailment makes it possible to formulate just one representative gold answer for each question, by using the gold answer as the hypothesis and the student answer as the text. This use of Textual Entailment can be understood as representing the possibilities for surface variation not explicitly in the "model", but implicitly in a more intelligent matching mechanism.

Next to ungrammatical input, a main challenge in this task is the length of the answers which often span considerably more than one sentence. Therefore, both text and hypothesis are generally decomposed into a set of atomic propositions, called "concepts" (Sukkarieh and Stoyanchev 2009) or "facets" (Nielsen et al. 2009). The role of a entailment engine is then specifically to determine pairs of facets where either (a) the student facet entails the gold facet without being entailed by a question facet (evidence for

a good, informative answer) or (b) the student facet and the gold facet are contradictory (evidence for a bad answer). Aggregate statistics over these two types of facet pairs, combined with information about facets in the gold answer that are not covered by the student answer (i.e., missing information) can then be used to compute an overall score for the student answer.

Machine Translation Evaluation. Evaluating the output of machine translation systems (cf. Chapter 32) is crucial for the progress of MT, yet human evaluation is costly and difficult to do reliably, which has led to the development of *automatic* measures of translation quality which score the system output with respect to a human-provided reference translation. Shallow methods are widely used for this purposes, such as BLEU score, which computes n-gram overlap. With the improving state-of-the-art in machine translation, however, studies such as Callison-Burch et al. (2006) have identified problems with shallow methods that mirror the problems of surface matching approaches to Textual Entailment. Consider the following two examples:

(26.7) *System:* This was declared terrorism by observers and witnesses.

Reference: Commentators as well as eyewitnesses are terming it terrorism.

(26.8) *System:* *BBC Haroon Rasheed Lal Masjid, Jamia Hafsa after* his *visit to* Auob Medical *Complex says Lal Masjid* and seminary in under a land mine.

Reference: What does BBC's Haroon Rasheed say after a visit to Lal Masjid Jamia Hafsa complex? There are no underground tunnels in Lal Masjid or Jamia Hafsa.

Example 26.7 on the previous page shows a good translation with a low BLEU score due to differences in word order as well as lexical choice. Example 26.8 on the preceding page shows a very bad translation which nevertheless receives a high BLEU score since almost all of the reference words appear almost literally in the hypothesis (marked in italics).

Padó et al. (2009) start out from the observation made in Section 26.1 on page 2 that a good translation *means* the same thing as the reference translation. They apply the Stanford Entailment Recognizer to pairs of system translations and reference translations, and compute entailment features for both directions. Then, they change the prediction stage from a binary prediction to a real-valued prediction using a linear regression model, which is trained on MT datasets with human judgments. The resulting metric correlates better to human judgments than surface matching-based metrics. For Example 26.7 on the preceding page, the entailment system abstracts away from word order, determines that the two main verbs are paraphrases of one another, that the corresponding argument heads are synonyms (*it/this*, *commentator/observer*), and recognizes the equivalent realizations of the coordination (*and/as well as*), which leads to an overall good prediction.

For Example 26.8 on page 33, the system predicts a bad score based on a number of mismatch features that have fired. They indicate problems with the structural well-formedness of the MT output as well as semantic incompatibility between hypothesis and reference (argument structure and reference mismatches).

26.5.3 Entailment for Generation

A crucial component of state-of-the-art statistical machine translation systems is the translation model, a probability distribution over pairs of source and target language phrases. Generally, the precision of the translation increases with the length of the phrases, but sparsity increases as well. In particular, unknown words, which may occur frequently for domain-specific text or languages where few resources are available, cannot be translated at all and are usually omitted or copied verbatim. Mirkin et al. (2009b) generate alternatives for source language sentences that omit unknown words, and find consistent improvements both over a baseline without entailment and a paraphrase-based reformulation approach. This is an instance of using entailment to generate possible new hypotheses H for a given text T, an idea that might also be applicable, for example, to query expansion.

26.5.4 Entailment for Structuring Information

A final emerging application of Textual Entailment is to identify structure inherent in information expressed as a set of sentences, in order to provide a compact human-readable characterization of the information.

The first example is multi-document summarization, where the goal is

to compress a set of original texts into one short text which still contains as much information as possible (cf. Chapter 37 for details). Harabagiu et al. (2007) employ entailment in this task to validate candidate summaries created with more efficient, shallow matching methods. The validation first computes all pairwise entailments between sentences from the summaries to identify "semantic content units", i.e., clusters of sentences that correspond to distinct propositions. Then, the summaries are ranked by how well they cover the semantic content units. In this setup, the role of entailment is to encourage summaries to cover as much of the semantic content of the original texts as possible independently of how it is expressed.

Berant et al. (2010) consider all sentences that were identified as relevant for a query such as "What affects blood pressure?". In order to present this information succinctly, they define a graph whose nodes are predicates and whose edges indicate entailment relations, and learn an optimal set of edges. This set of edges can be interpreted as a hierarchical summary for a set of sentences relevant for the query.

26.6 Conclusions and Outlook

In this this chapter, we have described Textual Entailment, an applied framework for modelling inference relations in text. Textual Entailment can be seen as a platform for the evaluation of different semantic processing methods – or as a basis for "inference engines" that meet the semantic processing needs of diverse NLP applications.

Traditional work on formal semantics has usually treated inference as a

problem of normalization: once natural language sentences can be translated into a formal representation and combined with a knowledge base (such as description logics with an ontology, or a first- or higher-order logic with a full knowledge base), inference corresponds directly to reasoning in the logical calculus. However, all three steps – construction of logical representations, building of large, logically consistent knowledge bases and efficient reasoning – are difficult research questions that can be avoided by modeling inference directly on the natural language level, as Textual Entailment does.

To fulfill its goals, Textual Entailment needs to be consolidated in the shape of entailment engines if it wants to cash in on the promise of ready-to-go semantic processing that we envisaged in the Introduction. That is, we need specifications (or more concretely, APIs) both to the application side and the processing side: On the application side, NLP applications would ideally be able to use semantic processing with the same ease that they can use syntactic parsers, without being concerned about the inner workings of an entailment engine. On the processing side, there must be a well-defined, and ideally modular, architecture that defines reasoning mechanisms and knowledge representation and allows incremental, distributed system development and the creation of exchangeable, reusable knowledge and inference modules.

Further Reading and Relevant Resources

There are two other current articles that discuss Textual Entailment and its processing methods, Androutsopoulos and Malakasiotis (2010) and Sammons et al. (2012), as well as one book (Dagan et al. 2012).

More details on the Recognizing Textual Entailment Challenges (Section 26.2) can be found in the task overview papers (Dagan et al. 2005; Bar-Haim et al. 2006; Giampiccolo et al. 2007, 2008; Bentivogli et al. 2009b, 2010, 2011). Another source for papers on this topic is the set of proceedings of workshops on this topic sponsored by the Association of Computational Linguistics (ACL), including Dolan and Dagan (2005); Sekine and Inui (2007); Callison-Burch and Zanzotto (2009); Padó and Thater (2011). Two tutorials can be found at `http://u.cs.biu.ac.il/~dagan/TE-Tutorial-ACL07.ppt` (ACL 2007) and `http://l2r.cs.uiuc.edu/~cogcomp/presentations/RTE_NAACL_2010.zip` (NAACL 2010).

Regarding resources, the ACL wiki contains a portal and "resource pool" (data sets, knowledge resources etc.) for Textual Entailment at `http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Portal`. Some entailment engines that are publicly available are BIUTEE (`http://u.cs.biu.ac.il/~nlp/downloads/biutee/protected-biutee.html`), EDITS (`http://edits.fbk.eu/`), Nutcracker (`http://svn.ask.it.usyd.edu.au/trac/candc/wiki/nutcracker`). and VENSES (`http://project.cgm.unive.it/venses.html`). The EC-funded project EXCITEMENT (FP7, 2012-14, `http://www.project-excitement.eu`) has as one of its goals to provide a general platform for entailment engines that can accommodate arbitrary components (algorithms and resources) and provides reusability.

Acknowledgments. The authors acknowledge the support of the European Commission under the EXCITEMENT project (FP7-ICT-287923).

Notes

[1]See http://www.celect.it/resources.php?id_page=rte.

[2]See Giampiccolo et al. (2008) and http://www.nist.gov/tac/tracks/2008/rte/rte.08.guidelines.html.

[3]See http://u.cs.biu.ac.il/~nlp/downloads/index.htm.

Bibliography

Androutsopoulos, Ion and Prodromos Malakasiotis (2010). 'A Survey of Paraphrasing and Textual Entailment Methods'. *Journal of Artificial Intelligence Research*, 38, 135–187.

Bannard, Colin and Chris Callison-Burch (2005). 'Paraphrasing with bilingual parallel corpora'. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 597–604. Ann Arbor, Michigan.

Bar-Haim, Roy, Jonathan Berant, and Ido Dagan (2009). 'A compact forest for scalable inference over entailment and paraphrase rules'. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1056–1065. Singapore.

Bar-Haim, Roy, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor (2006). 'The second PASCAL Recognising Textual Entailment Challenge'. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*. Venice, Italy.

Bar-Haim, Roy, Ido Dagan, Iddo Greental, and Eyal Shnarch (2007). 'Se-

mantic inference at the lexical-syntactic level'. In *Proceedings of the 22nd Conference on Artificial Intelligence*, 871–876. Vancouver, BC.

Bar-Haim, Roy, Idan Szpektor, and Oren Glickman (2005). 'Definition and analysis of intermediate entailment levels'. In *Proceedings of the ACL-PASCAL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, 55–60. Ann Arbor, MI.

Ben Aharon, Roni, Idan Szpektor, and Ido Dagan (2010). 'Generating entailment rules from FrameNet'. In *Proceedings of the ACL 2010 Conference Short Papers*, 241–246. Association for Computational Linguistics, Uppsala, Sweden.

Bentivogli, Luisa, Peter Clark, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo (2010). 'The sixth PASCAL recognising textual entailment challenge'. In *Proceedings of the TAC 2010 Workshop on Textual Entailment*. Gaithersburg, MD.

Bentivogli, Luisa, Peter Clark, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo (2011). 'The seventh PASCAL recognising textual entailment challenge'. In *Proceedings of the TAC 2011 Workshop on Textual Entailment*. Gaithersburg, MD.

Bentivogli, Luisa, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, Medea Lo Leggio, and Bernardo Magnini (2009a). 'Considering discourse references in textual entailment annotation'. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon*. Pisa, Italy.

Bentivogli, Luisa, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo (2009b). 'The fifth PASCAL recognising textual entailment challenge'. In *Proceedings of the TAC 2009 Workshop on Textual Entailment*. Gaithersburg, MD.

Berant, Jonathan, Ido Dagan, and Jacob Goldberger (2010). 'Global learning of focused entailment graphs'. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1220–1229. Uppsala, Sweden.

Bergmair, Richard (2009). 'A proposal on evaluation measures for RTE'. In *Proceedings of the ACL Workshop on Applied Textual Inference*, 10–17. Singapore.

Bobrow, D. G., C. Condoravdi, R. Crouch, V. De Paiva, L. Karttunen, T. H. King, R. Nairn, L. Price, and A. Zaenen (2007). 'Precision-focused textual inference'. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 16–21. Prague, Czech Republic.

Bos, Johan and Katja Markert (2005). 'Recognising textual entailment with logical inference'. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 628–635. Vancouver, BC.

Bos, Johan, Marco Pennacchiotti, and Fabio M. Zanzotto (2009). 'Textual entailment at EVALITA 2009'. In *Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*. Reggio Emilia.

Burchardt, Aljoscha, Marco Pennacchiotti, Stefan Thater, and Manfred Pinkal (2009). 'Assessing the impact of frame semantics on textual entailment'. *Journal of Natural Language Engineering*, 15(4), 527–550.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn (2006). 'Re-evaluating the role of BLEU in machine translation research'. In *Proceedings of the 11th Conference of the European Chapter of the ACL*, 249–256. Trento, Italy.

Callison-Burch, Chris and Fabio Massimo Zanzotto, editors (2009). *Proceedings of the ACL 2009 workshop on Textual Inference*. Singapore.

Carletta, Jean C. (1996). 'Assessing agreement on classification tasks: the kappa statistic'. *Computational Linguistics*, 22(2), 249–254.

Celikyilmaz, Asli, Marcus Thint, and Zhiheng Huang (2009). 'A graph-based semi-supervised learning for question-answering'. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 719–727. Singapore.

Chierchia, Gennaro and Sally McConnell-Ginet (2001). *Meaning and grammar: An introduction to semantics*. MIT Press, Cambridge, MA, 2nd edition.

Chklovski, Timothy and Patrick Pantel (2004). 'Verbocean: Mining the web for fine-grained semantic verb relations'. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 33–40. Barcelona, Spain.

Clark, P., W. R. Murray, J. Thompson, P. Harrison, J. Hobbs, and C. Fellbaum (2007). 'On the role of lexical and world knowledge in RTE-3'. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 54–59.

Clark, Peter, Christiane Fellbaum, Jerry R. Hobbs, Phil Harrison, William R. Murray, and John Thompson (2008). 'Augmenting WordNet for Deep Understanding of Text'. In *Proceedings of the Conference on Semantics in Text Processing*, 45–57. Venice, Italy.

Dagan, Ido, Bill Dolan, Bernardo Magnini, and Dan Roth (2009). 'Recognizing textual entailment: Rational, evaluation and approaches'. *Journal of Natural Language Engineering*, 15(4), i–xvii.

Dagan, Ido and Oren Glickman (2004). 'Probabilistic textual entailment: Generic applied modeling of language variability'. In *PASCAL workshop on Learning Methods for Text Understanding and Mining*. Grenoble, France.

Dagan, Ido, Oren Glickman, and Bernardo Magnini (2005). 'The PASCAL Recognising Textual Entailment Challenge'. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment*. Southampton, UK.

Dagan, Ido, Dan Roth, and Fabio Massimo Zanzotto (2012). *Recognizing Textual Entailment: Models and Applications*. Number 17 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool. In Press.

Danescu-Niculescu-Mizil, Cristian and Lillian Lee (2010). 'Don't 'Have a Clue'? Unsupervised Co-Learning of Downward-Entailing Operators'. In

Proceedings of the ACL 2010 Conference Short Papers, 247–252. Uppsala, Sweden.

De Marneffe, Marie-Catherine, Anna N. Rafferty, and Christopher D. Manning (2008). 'Finding contradictions in text'. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, 1039–1047. Columbus, OH.

de Salvo Braz, Ricardo, Roxana Girju, Vasin Punyakanok, Dan Roth, and Mark Sammons (2005). 'An inference model for semantic entailment in natural language'. In *Proceedings of the National Conference on Artificial Intelligence*, 1678–1679. Pittsburgh, PA.

Dinu, Georgiana and Rui Wang (2009). 'Inference rules and their application to recognizing textual entailment'. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, 211–219. Athens, Greece.

Dolan, Bill and Ido Dagan, editors (2005). *Proceedings of the ACL 2005 Workshop on the Empirical Modeling of Semantic Equivalence and Entailment*. Ann Arbor, MI.

Fellbaum, Christiane, editor (1998). *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London. ISBN 978-0-262-06197-1.

Fillmore, Charles J., Christopher R. Johnson, and Miriam R.L. Petruck (2003). 'Background to FrameNet'. *International Journal of Lexicography*, 16, 235–250.

Garoufi, Konstantina (2008). *Towards a Better Understanding of Applied Textual Entailment: Annotation and Evaluation of the RTE-2 Dataset*. Master's thesis, Saarland University.

Giampiccolo, Danilo, Bernardo Magnini, Ido Dagan, and Bill Dolan (2007). 'The third PASCAL recognising textual entailment challenge'. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Prague, Czech Republic.

Giampiccolo, Danilo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, and Elena Cabrio (2008). 'The fourth PASCAL recognising textual entailment challenge'. In *Proceedings of the TAC 2008 Workshop on Textual Entailment*. Gaithersburg, MD.

Harabagiu, Sanda and Andrew Hickl (2006). 'Methods for using textual entailment in open-domain question answering'. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 905–912. Sydney, Australia.

Harabagiu, Sanda, Andrew Hickl, and Finley Lacatusu (2007). 'Satisfying information needs with multi-document summaries'. *Information Processing and Management*, 43(6), 1619–1642. ISSN 0306-4573.

Harmeling, Stefan (2009). 'Inferring textual entailment with a probabilistically sound calculus'. *Journal of Natural Language Engineering*, 459–477.

Heilman, Michael and Noah A. Smith (2010). 'Tree edit models for recognizing textual entailments, paraphrases, and answers to questions'. In *Human*

Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 1011–1019. Los Angeles, CA.

Hickl, Andrew and Jeremy Bensley (2007). 'A discourse commitment-based framework for recognizing textual entailment'. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 171–176. Prague, Czech Republic.

Hickl, Andrew, John Williams, Jeremy Bensley, Kirk Roberts, Bryan Rink, and Ying Shi (2006). 'Recognizing textual entailment with LCCs Groundhog system'. In *Proceedings of the Second PASCAL Challenges Workshop*. Venice, Italy.

Kay, Paul (1987). 'Three properties of the ideal reader'. In Freedle, Roy O. and Richard P. Duraán, editors, *Cognitive and Linguistic Analyses of Test Performance*, 208–224. Ablex, Norwood, NJ.

Kipper, Karin, Hoa Trang Dang, and Martha Palmer (2000). 'Class-based construction of a verb lexicon'. In *Proceedings of the National Conference on Artificial Intelligence*, 691–696. Austin, TX.

Kotlerman, Lili, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet (2010). 'Directional distributional similarity for lexical inference'. *Natural Language Engineering*, 16(4), 359–389. ISSN 1351-3249.

Lin, Dekang (1998a). 'Automatic retrieval and clustering of similar words'. In *Proceedings of the International Conference on Computational Linguistics*

and Annual Meeting of the Association for Computational Linguistics, 768–774. Montréal, Canada.

Lin, Dekang (1998b). 'A dependency-based method for evaluating broad-coverage parsers'. *Natural Language Engineering*, 4, 97–114.

Lin, Dekang and Patrick Pantel (2002). 'Discovery of inference rules for question answering'. *Journal of Natural Language Engineering*, 7(4), 343–360.

MacCartney, Bill, Michel Galley, and Christopher D. Manning (2008). 'A phrase-based alignment model for natural language inference'. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 802–811. Honolulu, Hawaii.

MacCartney, Bill, Trond Grenager, Marie-Catherine de Marneffe, Daniel Cer, and Christopher D. Manning (2006). 'Learning to recognize features of valid textual entailments'. In *Proceedings of the Human Language Technology Conference of the NAACL*, 41–48. New York City, USA.

MacCartney, Bill and Christopher D. Manning (2007). 'Natural logic for textual inference'. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 193–200. Prague, Czech Republic.

MacCartney, Bill and Christopher D. Manning (2008). 'Modeling semantic containment and exclusion in natural language inference'. In *Proceedings of the 22nd International Conference on Computational Linguistics*, 521–528. Manchester, UK.

Manning, Christopher D. (2006). 'Local Textual Inference: It's hard to circumscribe, but you know it when you see it - and NLP needs it'. `http://nlp.stanford.edu/~manning/papers/LocalTextualInference.pdf`.

Mirkin, Shachar, Ido Dagan, and Sebastian Padó (2010). 'Assessing the role of discourse references in entailment inference'. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1209–1219. Uppsala, Sweden.

Mirkin, Shachar, Ido Dagan, and Eyal Shnarch (2009a). 'Evaluating the inferential utility of lexical-semantic resources'. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, 558–566. Athens, Greece.

Mirkin, Shachar, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor (2009b). 'Source-language entailment modeling for translating unknown terms'. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 791–799. Singapore.

Moldovan, Dan and Vasile Rus (2001). 'Logic form transformation of wordnet and its applicability to question answering'. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, 402–409. Toulouse, France.

Monz, Christof and Maarten de Rijke (2001). 'Light-weight entailment checking for computational semantics'. In *Proceedings of the Conference on Inference in Computational Semantics*, 59–72. Siena, Italy.

Nairn, Rowan, Cleo Condoravdi, and Lauri Karttunen (2006). 'Computing relative polarity for textual inference'. In *In Proceedings of the Conference on Inference in Computational Semantics*. Buxton, England.

Negri, Matteo, Milen Kouylekov, Bernardo Magnini, Yashar Mehdad, and Elena Cabrio (2009). 'Towards extensible textual entailment engines: the EDITS package'. In *Proceeding of the 11th Conference of the Italian Association for Artificial Intelligence*. Reggio Emilia, Italy.

Nielsen, Rodney D., Wayne Ward, and James H. Martin (2009). 'Recognizing entailment in intelligent tutoring systems'. *Journal of Natural Language Engineering*, 15(4), 479–501.

Norvig, Peter (1983). 'Six problems for story understanders'. In *Proceedings of the National Conference on Artificial Intelligence*. Washington, DC.

Norvig, Peter (1987). 'Inference in text understanding'. In *Proceedings of the National Conference on Artificial Intelligence*, 561–565. Seattle, WA.

Padó, Sebastian, Michel Galley, Dan Jurafsky, and Christopher D. Manning (2009). 'Robust machine translation evaluation with entailment features'. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 297–305. Singapore.

Padó, Sebastian and Stefan Thater, editors (2011). *Proceedings of the EMNLP 2011 workshop on Textual Inference*. Edinburgh, UK.

Pantel, Patrick, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski,

and Eduard H. Hovy (2007). 'ISP: Learning inferential selectional preferences'. In *Proceedings of the Conference of the North American Chapter of the ACL*, 564–571. Rochester, NY.

Peñas, Anselmo, Álvaro Rodrigo, Valentín Sama, and Felisa Verdejo (2008). 'Testing the reasoning for question answering validation'. *Journal of Logic and Computation*, 18, 459–474.

Raina, Rajat, Andrew Y. Ng, and Christopher D. Manning (2005). 'Robust textual inference via learning and abductive reasoning'. In *Proceedings of the 20th Conference on Artificial Intelligence*, 1099–1105. Pittsburgh, PA.

Romano, Lorenza, Milen Kouylekov, Idan Szpektor, Ido Dagan, and Alberto Lavelli (2006). 'Investigating a generic paraphrase-based approach for relation extraction'. In *Proceedings of the 11th Conference of the European Chapter of the ACL*, 401–408. Trento, Italy.

Roth, D., M. Sammons, and V.G. Vydiswaran (2009). 'A framework for entailed relation recognition'. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, 57–60. Singapore.

Salton, Gerard and M. J. McGill (1983). *Introduction to Modern Information Retrieval*. McGraw Hill Book Co., New York.

Sammons, M., V. Vydiswaran, and D. Roth (2012). 'Recognizing textual entailment (forthcoming)'. In Bikel, Daniel M. and Imed Zitouni, editors, *Multilingual Natural Language Applications: From Theory to Practice*. Prentice Hall.

Sekine, Satoshi and Kentaro Inui, editors (2007). *Proceedings of the ACL 2007 workshop on Textual Inference and Paraphrasing*. Prague, Czech Republic.

Shnarch, Eyal, Libby Barak, and Ido Dagan (2009). 'Extracting Lexical Reference Rules from Wikipedia'. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 450–458. Singapore.

Shnarch, Eyal, Jacob Goldberger, and Ido Dagan (2011). 'A probabilistic modeling framework for lexical entailment'. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 558–563. Portland, OR.

Snow, Rion, Daniel Jurafsky, and Andrew Y. Ng (2006). 'Semantic taxonomy induction from heterogenous evidence'. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 801–808. Sydney, Australia.

Stern, Asher and Ido Dagan (2011). 'A confidence model for syntactically-motivated entailment proofs'. In *Proceedings of the Conference on Recent Advances in Natural Language Processing*, 455–462. Borovets, Bulgaria.

Stern, Asher and Ido Dagan (2012). 'A modular open-source system for recognizing textual entailment'. In *Proceedings the Demo Session of the 50th Annual Meeting of the ACL*. Jeju Island, South Korea.

Stern, Asher, Roni Stern, Ido Dagan, and Ariel Felner (2012). 'Efficient search for transformation-based inference'. In *Proceedings of the 50th Annual Meeting of the ACL*. Jeju Island, South Korea.

Sukkarieh, Jana and Svetlana Stoyanchev (2009). 'Automating model building in c-rater'. In *Proceedings of the ACL Workshop on Applied Textual Inference*, 61–69. Singapore.

Szpektor, Idan and Ido Dagan (2008). 'Learning Entailment Rules for Unary Templates'. In *Proceedings of the 22nd International Conference on Computational Linguistics*, 849–856. Manchester, UK.

Szpektor, Idan and Ido Dagan (2009). 'Augmenting wordnet-based inference with argument mapping'. In *Proceedings of the ACL Workshop on Applied Textual Inference*, 27–35. Singapore.

Szpektor, Idan, Ido Dagan, Roy Bar-Haim, and Jacob Goldberger (2008). 'Contextual preferences'. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, 683–691. Columbus, OH.

Szpektor, Idan, Hristo Tanev, Ido Dagan, and Bonaventura Coppola (2004). 'Scaling web-based acquisition of entailment relations'. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 41–48. Barcelona, Spain.

Tatu, Marta and Dan Moldovan (2005). 'A semantic approach to recognizing textual entailment'. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 371–378. Vancouver, BC.

Vanderwende, Lucy and Bill Dolan (2006). 'What syntax can contribute in the entailment task'. In *Machine Learning Challenges*, volume 2944 of *Lecture Notes in Computer Science*, 205–216. Springer.

Wang, Mengqiu and Christopher Manning (2010). 'Probabilistic tree-edit models with structured latent variables for textual entailment and question answering'. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 1164–1172. Beijing, China.

Zaenen, Annie, Richard Crouch, and Lauri Karttunen (2006). 'Circumscribing is not excluding: A reply to Manning'. `http://www2.parc.com/istl/members/karttune/publications/reply-to-manning.pdf`.

Zaenen, Annie, Lauri Karttunen, and Richard Crouch (2005). 'Local textual inference: Can it be defined or circumscribed?' In *Proceedings of the ACL-PASCAL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, 31–36.

Zanzotto, Fabio Massimo, Marco Pennacchiotti, and Alessandro Moschitti (2009). 'A machine learning approach to textual entailment recognition'. *Journal of Natural Language Engineering*, 15(4), 551–582.

Zhao, Shiqi, Haifeng Wang, Ting Liu, and Sheng Li (2008). 'Pivot approach for extracting paraphrase patterns from bilingual corpora'. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, 780–788. Columbus, OH.