Definition and Analysis of Intermediate Entailment Levels

Roy Bar-Haim, Idan Szpektor, Oren Glickman
Computer Science Department
Bar Ilan University
Ramat-Gan 52900, Israel
{barhair,szpekti,glikmao}@cs.biu.ac.il

Abstract

In this paper we define two intermediate models of textual entailment, which correspond to lexical and lexical-syntactic levels of representation. We manually annotated a sample from the RTE dataset according to each model, compared the outcome for the two models, and explored how well they approximate the notion of entailment. We show that the lexical-syntactic model outperforms the lexical model, mainly due to a much lower rate of false-positives, but both models fail to achieve high recall. Our analysis also shows that *paraphrases* stand out as a dominant contributor to the entailment task. We suggest that our models and annotation methods can serve as an evaluation scheme for entailment at these levels.

1 Introduction

Textual entailment has been proposed recently as a generic framework for modeling semantic variability in many Natural Language Processing applications, such as Question Answering, Information Extraction, Information Retrieval and Document Summarization. The textual entailment relationship holds between two text fragments, termed text and hypothesis, if the truth of the hypothesis can be inferred from the text.

Identifying entailment is a complex task that incorporates many levels of linguistic knowledge and inference. The complexity of modeling entailment was demonstrated in the first PASCAL Challenge Workshop on Recognizing Textual Entailment (RTE) (Dagan et al., 2005). Systems that participated in the challenge used various combinations of NLP components in order to perform entailment inferences. These components can largely be classified as operating at the lexical, syntactic and semantic levels (see Table 1 in (Dagan et al., 2005)). However, only little research was done to analyze the contribution of each inference level, and on the contribution of individual inference mechanisms within each level.

This paper suggests that decomposing the complex task of entailment into subtasks, and analyzing the contribution of individual NLP components for these subtasks would make a step towards better understanding of the problem, and for pursuing better entailment engines. We set three goals in this paper. First, we consider two modeling levels that employ only part of the inference mechanisms, but perform perfectly at each level. We explore how well these models approximate the notion of entailment, and analyze the differences between the outcome of the different levels. Second, for each of the presented levels, we evaluate the distribution (and contribution) of each of the inference mechanisms typically associated with that level. Finally, we suggest that the definitions of entailment at different levels of inference, as proposed in this paper, can serve as guidelines for manual annotation of a "gold standard" for evaluating systems that operate at a particular level. Altogether, we set forth a possible methodology for annotation and analysis of entail-

Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, pages 55–60,
Ann Arbor, June 2005.

ment datasets.

We introduce two levels of entailment: *Lexical* and *Lexical-Syntactic*. We propose these levels as intermediate stages towards a complete entailment model. We define an entailment model for each level and manually evaluate its performance over a sample from the RTE test-set. We focus on these two levels as they correspond to well-studied NLP tasks, for which robust tools and resources exist, e.g. parsers, part of speech taggers and lexicons. At each level we included inference types that represent common practice in the field. More advanced processing levels which involve logical/semantic inference are less mature and were left beyond the scope of this paper.

We found that the main difference between the lexical and lexical-syntactic levels is that the lexical-syntactic level corrects many false-positive inferences done at the lexical level, while introducing only a few false-positives of its own. As for identifying positive cases (recall), both systems exhibit similar performance, and were found to be complementary. Neither of the levels was able to identify more than half of the positive cases, which emphasizes the need for deeper levels of analysis. Among the different inference components, *paraphrases* stand out as a dominant contributor to the entailment task, while synonyms and derivational transformations were found to be the most frequent at the lexical level.

Using our definitions of entailment models as guidelines for manual annotation resulted in a high level of agreement between two annotators, suggesting that the proposed models are well-defined.

Our study follows on previous work (Vanderwende et al., 2005), which analyzed the RTE Challenge test-set to find the percentage of cases in which syntactic analysis alone (with optional use of thesaurus for the lexical level) suffices to decide whether or not entailment holds. Our study extends this work by considering a broader range of inference levels and inference mechanisms and providing a more detailed view. A fundamental difference between the two works is that while Vanderwende et al. did not make judgements on cases where additional knowledge was required beyond syntax, our entailment models were evaluated over all of the cases, including those that require higher levels of inference. This allows us to view the entailment model at each level as an idealized *system* approximating full entailment, and to evaluate its overall success.

The rest of the paper is organized as follows: section 2 provides definitions for the two entailment levels; section 3 describes the annotation experiment we performed, its results and analysis; section 4 concludes and presents planned future work.

2 Definition of Entailment Levels

In this section we present definitions for two entailment models that correspond to the *Lexical* and *Lexical-Syntactic* levels. For each level we describe the available inference mechanisms. Table 1 presents several examples from the RTE test-set together with annotation of entailment at the different levels.

2.1 The Lexical entailment level

At the lexical level we assume that the text T and hypothesis H are represented by a bag of (possibly multi-word) terms, ignoring function words. At this level we define that entailment holds between T and H if every term h in H can be matched by a corresponding entailing term t in T. t is considered as entailing h if either h and t share the same lemma and part of speech, or t can be matched with h through a sequence of lexical transformations of the types described below.

Morphological derivations This inference mechanism considers two terms as equivalent if one can be obtained from the other by some morphological derivation. Examples include nominalizations (e.g. 'acquisition $\Leftrightarrow$ acquire'), pertainyms (e.g. 'Afghanistan $\Leftrightarrow$ Afghan'), or nominal derivations like 'terrorist $\Leftrightarrow$ terror'.

Ontological relations This inference mechanism refers to ontological relations between terms. A term is inferred from another term if a chain of valid ontological relations between the two terms exists (Andreevskaia et al., 2005). In our experiment we regarded the following three ontological relations as providing entailment inferences: (1) 'synonyms' (e.g. 'free $\Leftrightarrow$ release' in example 1361, Table 1); (2) 'hypernym' (e.g. 'produce $\Rightarrow$ make') and (3) 'meronym-holonym' (e.g. 'executive $\Rightarrow$ company').

No.	Text	Hypothesis	Task	Ent.	Lex. Ent.	Syn. Ent.
322	Turnout for the historic vote for the first time since the EU took in 10 new members in May has hit a record low of 45.3%.	New members joined the EU.	IR	true	false	true
1361	A Filipino hostage in Iraq was released.	A Filipino hostage was freed in Iraq.	CD	true	true	true
1584	Although a Roscommon man by birth, born in Rooskey in 1932, Albert "The Slasher" Reynolds will forever be a Longford man by association.	Albert Reynolds was born in Co. Roscommon.	QA	true	true	true
1911	The SPD got just 21.5% of the vote in the European Parliament elections, while the conservative opposition parties polled 44.5%.	The SPD is defeated by the opposition parties.	IE	true	false	false
2127	Coyote shot after biting girl in Vanier Park.	Girl shot in park.	IR	false	true	false

Table 1: Examples of text-hypothesis pairs, taken from the PASCAL RTE test-set. Each line includes the example number at the RTE test-set, the text and hypothesis, the task within the test-set, whether entailment holds between the text and hypothesis (Ent.), whether Lexical entailment holds (Lex. Ent.) and whether Lexical-Syntactic entailment holds (Syn. Ent.).

Lexical World knowledge This inference mechanism refers to world knowledge reflected at the lexical level, by which the meaning of one term can be inferred from the other. It includes both knowledge about named entities, such as 'Taliban $\Rightarrow$ organization' and 'Roscommon $\Leftrightarrow$ Co. Roscommon' (example 1584 in Table 1), and other lexical relations between words, such as WordNet's relations 'cause' (e.g. 'kill $\Rightarrow$ die') and 'entail' (e.g. 'snore $\Rightarrow$ sleep').

2.2 The Lexical-syntactic entailment level

At the lexical-syntactic level we assume that the text and the hypothesis are represented by the set of syntactic dependency relations of their dependency parse. At this level we ignore determiners and auxiliary verbs, but do include relations involving other function words. We define that entailment holds between T and H if the relations within H can be "covered" by the relations in T. In the trivial case, lexical-syntactic entailment holds if all the relations composing H appear verbatim in T (while additional relations within T are allowed). Otherwise, such coverage can be obtained by a sequence of transformations applied to the relations in T, which should yield all the relations in H.

One type of such transformations are the lexical transformations, which replace corresponding lexical items, as described in sub-section 2.1. When applying morphological derivations it is assumed that the syntactic structure is appropriately adjusted. For example, "Mexico produces oil" can be mapped to "oil production by Mexico" (the NOMLEX resource (Macleod et al., 1998) provides a good example for systematic specification of such transformations).

Additional types of transformations at this level are specified below.

Syntactic transformations This inference mechanism refers to transformations between syntactic structures that involve the same lexical elements and preserve the meaning of the relationships between them (as analyzed in (Vanderwende et al., 2005)). Typical transformations include passive-active and apposition (e.g. 'An Wang, a native of Shanghai $\Leftrightarrow$ An Wang is a native of Shanghai').

Entailment paraphrases This inference mechanism refers to transformations that modify the syntactic structure of a text fragment as well as some of its lexical elements, while holding an entailment relationship between the original text and the transformed one. Such transformations are typically denoted as 'paraphrases' in the literature, where a wealth of methods for their automatic acquisition were proposed (Lin and Pantel, 2001; Shinyama et al., 2002; Barzilay and Lee, 2003; Szpektor et al., 2004). Following the same spirit, we focus here on transformations that are local in nature, which, according to the literature, may be amenable for large scale acquisition. Examples include: 'X is Y man by birth $\rightarrow$ X was born in Y' (example 1584 in Table 1), 'X take in Y $\Leftrightarrow$ Y join X'[1] and 'X is holy book of Y $\Rightarrow$ Y follow X'[2].

Co-reference Co-references provide equivalence relations between different terms in the text and thus induce transformations that replace one term in a text with any of its co-referenced terms. For example, the sentence "Italy and Germany have each played twice, and they haven't beaten anybody yet."[3] entails "Neither Italy nor Germany have won yet", involving the co-reference transformation 'they $\Rightarrow$ Italy and Germany'.

Example 1584 in Table 1 demonstrates the need to combine different inference mechanisms to achieve lexical-syntactic entailment, requiring world-knowledge, paraphrases and syntactic transformations.

3 Empirical Analysis

In this section we present the experiment that we conducted in order to analyze the two entailment levels, which are presented in section 2, in terms of relative performance and correlation with the notion of textual entailment.

3.1 Data and annotation procedure

The RTE test-set[4] contains 800 Text-Hypothesis pairs (usually single sentences), which are typical to various NLP applications. Each pair is annotated with a boolean value, indicating whether the hypothesis is entailed by the text or not, and the test-set is balanced in terms of positive and negative cases. We shall henceforth refer to this annotation as the *gold standard*. We constructed a sample of 240 pairs from four different tasks in the test-set, which correspond to the main applications that may benefit from entailment: information extraction (IE), information retrieval (IR), question answering (QA), and comparable documents (CD). We randomly picked 60 pairs from each task, and in total 118 of the cases were positive and 122 were negative.

In our experiment, two of the authors annotated, for each of the two levels, whether or not entailment can be established in each of the 240 pairs. The annotators agreed on 89.6% of the cases at the lexical level, and 88.8% of the cases at the lexical-syntactic level, with Kappa statistics of 0.78 and 0.73, respectively, corresponding to 'substantial agreement' (Landis and Koch, 1977). This relatively high level of agreement suggests that the notion of lexical and lexical-syntactic entailment we propose are indeed well-defined.

Finally, in order to establish statistics from the annotations, the annotators discussed all the examples they disagreed on and produced a final joint decision.

3.2 Evaluating the different levels of entailment

	L	LS
True positive (118)	52	59
False positive (122)	36	10
Recall	44%	50%
Precision	59%	86%
F_1	0.5	0.63
Accuracy	58%	71%

Table 2: Results per level of entailment.

Table 2 summarizes the results obtained from our annotated dataset for both lexical (L) and lexical-syntactic (LS) levels. Taking a "system"-oriented perspective, the annotations at each level can be viewed as the classifications made by an idealized system that includes a perfect implementation of the inference mechanisms in that level. The first two

[1] Example no 322 in the PASCAL RTE test-set.

[2] Example no 1575 in the PASCAL RTE test-set.

[3] Example no 298 in the PASCAL RTE test-set.

[4] The complete RTE dataset can be obtained at http://www.pascal-network.org/Challenges/RTE/Datasets/

rows show for each level how the cases, which were recognized as positive by this level (i.e. the entailment holds), are distributed between "true positive" (i.e. positive according to the gold standard) and "false positive" (negative according to the gold standard). The total number of positive and negative pairs in the dataset is reported in parentheses. The rest of the table details recall, precision, F_1 and accuracy.

The distribution of the examples in the RTE test-set cannot be considered representative of a real-world distribution (especially because of the controlled balance between positive and negative examples). Thus, our statistics are not appropriate for accurate prediction of application performance. Instead, we analyze how well these simplified models of entailment succeed in approximating "real" entailment, and how they compare with each other.

The proportion between true and false positive cases at the lexical level indicates that the correlation between lexical match and entailment is quite low, reflected in the low precision achieved by this level (only 59%). This result can be partly attributed to the idiosyncracies of the RTE test-set: as reported in (Dagan et al., 2005), samples with high lexical match were found to be biased towards the negative side. Interestingly, our measured accuracy correlates well with the performance of systems at the PASCAL RTE Workshop, where the highest reported accuracy of a lexical system is 0.586 (Dagan et al., 2005).

As one can expect, adding syntax considerably reduces the number of false positives - from 36 to only 10. Surprisingly, at the same time the number of true positive cases grows from 52 to 59, and correspondingly, precision rise to 86%. Interestingly, neither the lexical nor the lexical-syntactic level are able to cover more than half of the positive cases (e.g. example 1911 in Table 1).

In order to better understand the differences between the two levels, we next analyze the overlap between them, presented in Table 3. Looking at Table 3(a), which contains only the positive cases, we see that many examples were recognized only by one of the levels. This interesting phenomenon can be explained on the one hand by lexical matches that could not be validated in the syntactic level, and on the other hand by the use of paraphrases, which are introduced only in the lexical-syntactic level. (e.g. example 322 in Table 1).

		Lexical-Syntactic	
		H $\Rightarrow$ T	H $\not\Rightarrow$ T
Lexical	H $\Rightarrow$ T	38	14
	H $\not\Rightarrow$ T	21	45

(a) positive examples

		Lexical-Syntactic	
		H $\Rightarrow$ T	H $\not\Rightarrow$ T
Lexical	H $\Rightarrow$ T	7	29
	H $\not\Rightarrow$ T	3	83

(b) negative examples

Table 3: Correlation between the entailment levels. (a) includes only the positive examples from the RTE dataset sample, and (b) includes only the negative examples.

This relatively symmetric situation changes as we move to the negative cases, as shown in Table 3(b). By adding syntactic constraints, the lexical-syntactic level was able to fix 29 false positive errors, misclassified at the lexical level (as demonstrated in example 2127, Table 1), while introducing only 3 new false-positive errors. This exemplifies the importance of syntactic matching for precision.

3.3 The contribution of various inference mechanisms

Inference Mechanism	f	$\triangle$R	%
Synonym	19	14.4%	16.1%
Morphological	16	10.1%	13.5%
Lexical World knowledge	12	8.4%	10.1%
Hypernym	7	4.2%	5.9%
Mernoym	1	0.8%	0.8%
Entailment Paraphrases	37	26.2%	31.3%
Syntactic transformations	22	16.9%	18.6%
Coreference	10	5.0%	8.4%

Table 4: The frequency (f), contribution to recall ($\triangle$R) and percentage (%), within the gold standard positive examples, of the various inference mechanisms at each level, ordered by their significance.

In order to get a sense of the contribution of the various components at each level, statistics on the inference mechanisms that contributed to the coverage of the hypothesis by the text (either full or partial) were recorded by one annotator. Only the positive cases in the gold standard were considered.

For each inference mechanism we measured its frequency, its contribution to the recall of the related level and the percentage of cases in which it is required for establishing entailment. The latter also takes into account cases where only partial coverage could be achieved, and thus indicates the significance of each inference mechanism for any entailment system, regardless of the models presented in this paper. The results are summarized in Table 4.

From Table 4 it stands that paraphrases are the most notable contributors to recall. This result indicates the importance of paraphrases to the entailment task and the need for large-scale paraphrase collections. Syntactic transformations are also shown to contribute considerably, indicating the need for collections of syntactic transformations as well. In that perspective, we propose our annotation framework as means for evaluating collections of paraphrases or syntactic transformations in terms of recall.

Finally, we note that the co-reference moderate contribution can be partly attributed to the idiosyncracies of the RTE test-set: the annotators were guided to replace anaphors with the appropriate reference, as reported in (Dagan et al., 2005).

4 Conclusions

In this paper we presented the definition of two entailment models, Lexical and Lexical-Syntactic, and analyzed their performance manually. Our experiment shows that the lexical-syntactic level outperforms the lexical level in all measured aspects. Furthermore, paraphrases and syntactic transformations emerged as the main contributors to recall. These results suggest that a lexical-syntactic framework is a promising step towards a complete entailment model.

Beyond these empirical findings we suggest that the presented methodology can be used generically to annotate and analyze entailment datasets.

In future work, it would be interesting to analyze higher levels of entailment, such as logical inference and deep semantic understanding of the text.

Acknowledgements

We would like to thank Ido Dagan for helpful discussions and for his scientific supervision. This work was supported in part by the IST Programme of the European Community, under the *PASCAL Network of Excellence*, IST-2002-506778. This publication only reflects the authors' views.

References

Alina Andreevskaia, Zhuoyan Li and Sabine Bergler. 2005. Can Shallow Predicate Argument Structures Determine Entailment?. In *Proceedings of Pascal Challenge Workshop on Recognizing Textual Entailment, 2005*.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL 2003*. pages 16-23, Edmonton, Canada.

Ido Dagan, Bernardo Magnini and Oren Glickman. 2005. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of Pascal Challenge Workshop on Recognizing Textual Entailment, 2005*.

J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159-174.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for Question Answering. *Natural Language Engineering*, 7(4):343-360.

C. Macleod, R. Grishman, A. Meyers, L. Barrett and R. Reeves. 1998. Nomlex: A lexicon of nominalizations. In *Proceedings of the 8th International Congress of the European Association for Lexicography, 1998*. Liège, Belgium: EURALEX.

Yusuke Shinyama and Satoshi Sekine, Kiyoshi Sudo and Ralph Grishman. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of Human Language Technology Conference (HLT 2002)*. San Diego, USA.

Idan Szpektor, Hristo Tanev, Ido Dagan and Bonnaventura Coppola. 2004. Scaling Web-based Acquistion of Entailment Relations. In *Proceedings of EMNLP 2004*.

Lucy Vanderwende, Deborah Coughlin and Bill Dolan. 2005. What Syntax Contribute in Entailment Task. In *Proceedings of Pascal Challenge Workshop on Recognizing Textual Entailment, 2005*.