Recognizing Partial Textual Entailment

Omer Levy† **Torsten Zesch**§ **Ido Dagan**† **Iryna Gurevych**§

† Natural Language Processing Lab
Computer Science Department
Bar-Ilan University

§ Ubiquitous Knowledge Processing Lab
Computer Science Department
Technische Universität Darmstadt

Abstract

Textual entailment is an asymmetric relation between two text fragments that describes whether one fragment can be inferred from the other. It thus cannot capture the notion that the target fragment is "almost entailed" by the given text. The recently suggested idea of *partial textual entailment* may remedy this problem. We investigate partial entailment under the faceted entailment model and the possibility of adapting existing textual entailment methods to this setting. Indeed, our results show that these methods are useful for recognizing partial entailment. We also provide a preliminary assessment of how partial entailment may be used for recognizing (complete) textual entailment.

1 Introduction

Approaches for applied semantic inference over texts gained growing attention in recent years, largely triggered by the *textual entailment* framework (Dagan et al., 2009). Textual entailment is a generic paradigm for semantic inference, where the objective is to recognize whether a textual hypothesis (labeled H) can be inferred from another given text (labeled T). The definition of textual entailment is in some sense strict, in that it requires that H's meaning be implied by T *in its entirety*. This means that from an entailment perspective, a text that contains the main ideas of a hypothesis, but lacks a minor detail, is indiscernible from an entirely unrelated text. For example, if T is "muscles move bones", and H "the main job of muscles is to move bones", then T does *not* entail H, and we are left with no sense of how *close* (T, H) were to entailment.

In the related problem of semantic text similarity, gradual measures are already in use. The semantic text similarity challenge in SemEval 2012 (Agirre et al., 2012) explicitly defined different levels of similarity from 5 (semantic equivalence) to 0 (no relation). For instance, 4 was defined as "the two sentences are mostly equivalent, but some unimportant details differ", and 3 meant that "the two sentences are roughly equivalent, but some important information differs". Though this modeling does indeed provide finer-grained notions of similarity, it is not appropriate for semantic inference for two reasons. First, the term "important information" is vague; what makes one detail more important than another? Secondly, similarity is not sufficiently well-defined for sound semantic inference; for example, "snowdrops bloom in summer" and "snowdrops bloom in winter" may be similar, but have contradictory meanings. All in all, these measures of similarity do not quite capture the gradual relation needed for semantic inference.

An appealing approach to dealing with the rigidity of textual entailment, while preserving the more precise nature of the entailment definition, is by breaking down the hypothesis into components, and attempting to recognize whether each one is individually entailed by T. It is called *partial textual entailment*, because we are only interested in recognizing whether a single element of the hypothesis is entailed. To differentiate the two tasks, we will refer to the original textual entailment task as *complete textual entailment*.

Partial textual entailment was first introduced by Nielsen et al. (2009), who presented a machine learning approach and showed significant improvement over baseline methods. Recently, a public benchmark has become available through the Joint Student Response Analysis and 8th Recognizing Textual Entailment (RTE) Challenge in SemEval 2013 (Dzikovska et al., 2013), on which we focus in this paper.

Our goal in this paper is to investigate the idea of partial textual entailment, and assess whether

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 451–455,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

existing complete textual entailment methods can be used to recognize it. We assume the facet model presented in SemEval 2013, and adapt existing technologies to the task of recognizing partial entailment (Section 3). Our work further expands upon (Nielsen et al., 2009) by evaluating these adapted methods on the new RTE-8 benchmark (Section 4). Partial entailment may also facilitate an alternative divide and conquer approach to complete textual entailment. We provide an initial investigation of this approach (Section 5).

2 Task Definition

In order to tackle partial entailment, we need to find a way to decompose a hypothesis. Nielsen et al. (2009) defined a model of *facets*, where each such facet is a pair of words in the hypothesis and the direct semantic relation connecting those two words. We assume the simplified model that was used in RTE-8, where the relation between the words is not explicitly stated. Instead, it remains unstated, but its interpreted meaning should correspond to the manner in which the words are related in the hypothesis. For example, in the sentence "the main job of muscles is to move bones", the pair *(muscles, move)* represents a facet. While it is not explicitly stated, reading the original sentence indicates that *muscles* is the agent of *move*.

Formally, the task of recognizing faceted entailment is a binary classification task. Given a text T, a hypothesis H, and a facet within the hypothesis (w_1, w_2), determine whether the facet is either *expressed* or *unaddressed* by the text. Nielsen et al included additional classes such as *contradicting*, but in the scope of this paper we will only tend to the binary case, as was done in RTE-8.

Consider the following example:

T: Muscles generate movement in the body.
H: The main job of muscles is to move bones.

The facet *(muscles, move)* refers to the agent role in H, and is expressed by T. However, the facet *(move, bones)*, which refers to a theme or direct object relation in H, is unaddressed by T.

3 Recognizing Faceted Entailment

Our goal is to investigate whether existing entailment recognition approaches can be adapted to recognize faceted entailment. Hence, we specified relatively simple decision mechanisms over a set of entailment detection modules. Given a text and a facet, each module reports whether it recognizes entailment, and the decision mechanism then determines the binary class (*expressed* or *unaddressed*) accordingly.

3.1 Entailment Modules

Current textual entailment systems operate across different linguistic levels, mainly on lexical inference and syntax. We examined three representative modules that reflect these levels: *Exact Match*, *Lexical Inference*, and *Syntactic Inference*.

Exact Match We represent T as a bag-of-words containing all tokens and lemmas appearing in the text. We then check whether both facet lemmas w_1, w_2 appear in the text's bag-of-words. Exact matching was used as a baseline in previous recognizing textual entailment challenges (Bentivogli et al., 2011), and similar methods of lemma-matching were used as a component in recognizing textual entailment systems (Clark and Harrison, 2010; Shnarch et al., 2011).

Lexical Inference This feature checks whether both facet words, or semantically related words, appear in T. We use WordNet (Fellbaum, 1998) with the Resnik similarity measure (Resnik, 1995) and count a facet term w_i as matched if the similarity score exceeds a certain threshold (0.9, empirically determined on the training set). Both w_1 and w_2 must match for this module's entailment decision to be positive.

Syntactic Inference This module builds upon the open source[1] Bar-Ilan University Textual Entailment Engine (BIUTEE) (Stern and Dagan, 2011). BIUTEE operates on dependency trees by applying a sequence of knowledge-based transformations that converts T into H. It determines entailment according to the "cost" of generating the hypothesis from the text. The cost model can be automatically tuned with a relatively small training set. BIUTEE has shown state-of-the-art performance on previous recognizing textual entailment challenges (Stern and Dagan, 2012).

Since BIUTEE processes dependency trees, both T and the facet must be parsed. We therefore extract a path in H's dependency tree that represents the facet. This is done by first parsing H, and then locating the two nodes whose words compose the facet. We then find their lowest common ancestor (LCA), and extract the path P from w_1 to

[1] cs.biu.ac.il/~nlp/downloads/biutee

w_2 through the LCA. This path is in fact a dependency tree. BIUTEE can now be given T and P (as the hypothesis), and try to recognize whether the former entails the latter.

3.2 Decision Mechanisms

We started our experimentation process by defining *Exact Match* as a baseline. Though very simple, this unsupervised baseline performed surprisingly well, with 0.96 precision and 0.32 recall on expressed facets of the training data. Given its very high precision, we decided to use this module as an initial filter, and employ the others for classifying the "harder" cases.

We present all the mechanisms that we tested:

Baseline *Exact*
BaseLex *Exact* $\vee$ *Lexical*
BaseSyn *Exact* $\vee$ *Syntactic*
Disjunction *Exact* $\vee$ *Lexical* $\vee$ *Syntactic*
Majority *Exact* $\vee$ (*Lexical* $\wedge$ *Syntactic*)

Note that since every facet that *Exact Match* classifies as *expressed* is also *expressed* by *Lexical Inference*, *BaseLex* is essentially *Lexical Inference* on its own, and *Majority* is equivalent to the majority rule on all three modules.

4 Empirical Evaluation

4.1 Dataset: Student Response Analysis

We evaluated our methods as part of RTE-8. The challenge focuses on the domain of scholastic quizzes, and attempts to emulate the meticulous marking process that teachers do on a daily basis. Given a question, a student's response, and a reference answer, the task of *student response analysis* is to determine whether the student answered correctly. This task can be approximated as a special case of textual entailment; by assigning the student's answer as T and the reference answer as H, we are basically asking whether one can infer the correct (reference) answer from the student's response.

Recall the example from Section 2. In this case, H is a reference answer to the question:

Q: What is the main job of muscles?

T is essentially the student answer, though it is also possible to define T as the union of both the question and the student answer. In this work, we chose to exclude the question.

There were two tracks in the challenge: complete textual entailment (the main task) and partial entailment (the pilot task). Both tasks made use of the SciEntsBank corpus (Dzikovska et al., 2012), which is annotated at facet-level, and provides a convenient test-bed for evaluation of both partial and complete entailment. This dataset was split into train and test subsets. The test set has 16,263 facet-response pairs based on 5,106 student responses over 15 domains (learning modules). Performance was measured using micro-averaged F_1, over three different scenarios:

	Unseen Answers	Unseen Questions	Unseen Domains
Baseline	.670	.688	.731
BaseLex	.756	.710	.760
BaseSyn	.744	.733	.770
Disjunction	.695	.655	.703
Majority	**.782**	**.765**	**.816**

Table 1: Micro-averaged F_1 on the faceted SciEntsBank test set.

Unseen Answers Classify new answers to questions seen in training. Contains 464 student responses.

Unseen Questions Classify new answers to questions that were not seen in training, but other questions from the same domain were. Contains 631 student responses.

Unseen Domains Classify new answers to unseen questions from unseen domains. Contains 4,011 student responses.

4.2 Results

Table 1 shows the F_1-measure of each configuration in each scenario. There is some variance between the different scenarios; this may be attributed to the fact that there are much fewer *Unseen Answers* and *Unseen Questions* instances. In all cases, *Majority* significantly outperformed the other configurations. While *BaseLex* and *BaseSyn* improve upon the baseline, they seem to make different mistakes, in particular false positives. Their conjunction is thus a more conservative indicator of entailment, and proves helpful in terms of F_1. All improvements over the baseline were found to be statistically significant using McNemar's test with $p < 0.01$ (excluding *Disjunction*). It is also interesting to note that the systems' performance does not degrade in "harder" scenarios; this is a result of the mostly unsupervised nature of our modules.

Unfortunately, our system was the only submission in the partial entailment pilot track of RTE-8, so we have no comparisons with other systems. However, the absolute improvement from the exact-match baseline to the more sophisticated *Majority* is in the same ballpark as that of the best systems in previous recognizing textual entailment challenges. For instance, in the previous recognizing textual entailment challenge (Bentivogli et al., 2011), the best system yielded an F_1 score of 0.48, while the baseline scored 0.374. We can therefore conclude that existing approaches for recognizing textual entailment can indeed be adapted for recognizing partial entailment.

5 Utilizing Partial Entailment for Recognizing Complete Entailment

Encouraged by our results, we ask whether the same algorithms that performed well on the faceted entailment task can be used for recognizing complete textual entailment. We performed an initial experiment that examines this concept and sheds some light on the potential role of partial entailment as a possible facilitator for complete entailment.

We suggest the following 3-stage architecture:

1. Decompose the hypothesis into facets.
2. Determine whether each facet is entailed.
3. Aggregate the individual facet results and decide on complete entailment accordingly.

Facet Decomposition For this initial investigation, we use the facets provided in SciEntsBank; i.e. we assume that the step of *facet decomposition* has already been carried out. When the dataset was created for RTE-8, many facets were extracted automatically, but only a subset was selected. The facet selection process was done manually, as part of the dataset's annotation. For example, in "the main job of muscles is to move bones", the facet *(job, muscles)* was not selected, because it was not critical for answering the question. We refer to the issue of relying on manual input further below.

Recognizing Faceted Entailment This step was carried out as explained in the previous sections. We used the *Baseline* configuration and *Majority*, which performed best in our experiments above. In addition, we introduce *GoldBased* that uses the gold annotation of faceted entailment, and thus provides a certain upper bound on the performance of determining complete entailment based on facets.

	Unseen Answers	Unseen Questions	Unseen Domains
Baseline	.575	.582	.683
Majority	.707	.673	.764
GoldBased	.842	.897	.852
BestComplete	.773	.745	.712

Table 2: Micro-averaged F_1 on the 2-way complete entailment SciEntsBank test set.

Aggregation We chose the simplest sensible aggregation rule to decide on overall entailment: a student answer is classified as *correct* (i.e. it entails the reference answer) if it expresses each of the reference answer's facets. Although this heuristic is logical from a strict entailment perspective, it might yield false negatives on this particular dataset. This happens because tutors may sometimes grade answers as valid even if they omit some less important, or indirectly implied, facets.

Table 2 shows the experiment's results. The first thing to notice is that *GoldBased* is not perfect. There are two reasons for this behavior. First, the task of student response analysis is only an approximation of textual entailment, albeit a good one. This discrepancy was also observed by the RTE-8 challenge organizers (Dzikovska et al., 2013). The second reason is because some of the original facets were filtered when creating the dataset. This caused both false positives (when important facets were filtered out) and false negatives (when unimportant facets were retained).

Our *Majority* mechanism, which requires that the two underlying methods for partial entailment detection (*Lexical Inference* and *Syntactic Inference*) agree on a positive classification, bridges about half the gap from the baseline to the gold based method. As a rough point of comparison, we also show the performance of *BestComplete*, the winning entry in each setting of the RTE-8 main task. This measure is not directly comparable to our facet-based systems, because it did not rely on manually selected facets, and due to some variations in the dataset size (about 20% of the student responses were not included in the pilot task dataset). However, these results may indicate the

prospects of using faceted entailment for complete entailment recognition, suggesting it as an attractive research direction.

6 Conclusion and Future Work

In this paper, we presented an empirical attempt to tackle the problem of partial textual entailment. We demonstrated that existing methods for recognizing (complete) textual entailment can be successfully adapted to this setting. Our experiments showed that boolean combinations of these methods yield good results. Future research may add additional features and more complex feature combination methods, such as weighted sums tuned by machine learning. Furthermore, our work focused on a specific decomposition model – faceted entailment. Other flavors of partial entailment should be investigated as well. Finally, we examined the possibility of utilizing partial entailment for recognizing complete entailment in a semi-automatic setting, which relied on the manual facet annotation in the RTE-8 dataset. Our preliminary results suggest that this approach is indeed feasible, and warrant further research on facet-based entailment methods that rely on fully-automatic facet extraction.

Acknowledgements

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287923 (EXCITEMENT). We would like to thank the Minerva Foundation for facilitating this cooperation with a short term research grant.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A pilot on semantic textual similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation, in conjunction with the 1st Joint Conference on Lexical and Computational Semantics*, pages 385–393, Montreal, Canada.

Luisa Bentivogli, Peter Clark, Ido Dagan, Hoa Dang, and Danilo Giampiccolo. 2011. The seventh pascal recognizing textual entailment challenge. *Proceedings of TAC*.

Peter Clark and Phil Harrison. 2010. Blue-lite: a knowledge-based lexical entailment system for rte6. *Proc. of TAC*.

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

Myroslava O Dzikovska, Rodney D Nielsen, and Chris Brew. 2012. Towards effective tutorial feedback for explanation questions: A dataset and baselines. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 200–210. Association for Computational Linguistics.

Myroslava O. Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In **SEM 2013: The First Joint Conference on Lexical and Computational Semantics*, Atlanta, Georgia, USA, 13-14 June. Association for Computational Linguistics.

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

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

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 1995)*, pages 448–453.

Eyal Shnarch, 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*, pages 558–563, Portland, Oregon, USA, June. Association for Computational Linguistics.

Asher Stern and Ido Dagan. 2011. A confidence model for syntactically-motivated entailment proofs. In *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011)*, pages 455–462.

Asher Stern and Ido Dagan. 2012. Biutee: A modular open-source system for recognizing textual entailment. In *Proceedings of the ACL 2012 System Demonstrations*, pages 73–78, Jeju Island, Korea, July. Association for Computational Linguistics.