UIO-Lien: Entailment Recognition using Minimal Recursion Semantics

Elisabeth Lien
Department of Informatics
University of Oslo, Norway
`elien@ifi.uio.no`

Milen Kouylekov
Department of Informatics
University of Oslo, Norway
`milen@ifi.uio.no`

Abstract

In this paper we present our participation in the Semeval 2014 task *"Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment"*. Our results demonstrate that using generic tools for semantic analysis is a viable option for a system that recognizes textual entailment. The invested effort in developing such tools allows us to build systems for reasoning that do not require training.

1 Introduction

Recognizing textual entailment (RTE) has been a popular area of research in the last years. It has appeared in a variety of evaluation campaigns as both monolingual and multilingual tasks. A wide variety of techniques based on different levels of text interpretation has been used, e.g., lexical distance, dependency parsing and semantic role labeling (Androutsopoulos and Malakasiotis, 2010).

Our approach uses a semantic representation formalism called Minimal Recursion Semantics (MRS), which, to our knowledge, has not been used extensively in entailment decision systems. Notable examples of systems that use MRS are Wotzlaw and Coote (2013), and Bergmair (2010). In Wotzlaw and Coote (2013), the authors present an entailment recognition system which combines high-coverage syntactic and semantic text analysis with logical inference supported by relevant background knowledge. MRS is used as an intermediate format in transforming the results of the linguistic analysis into representations used for logical reasoning. The approach in Bergmair (2010) uses the syllogism as an approximation of natural language reasoning. MRS is used as a step in the translation of natural language sentences into logical formulae that are suitable for processing. Both works describe approaches that can be adapted to RTE, but no empirical evaluation is included to demonstrate the potential of the proposed approaches.

In contrast to these approaches, our system bases entailment decision directly on the MRS representations. Graph alignment over MRS representations forms the basis for entailment recognition. If key nodes in the hypothesis MRS can be aligned to nodes in the text MRS, this is treated as an indicator of entailment.

This paper represents our first attempt to evaluate a system based on logical-form semantic representations in a RTE competition. Using a state-of-the-art semantic analysis component, we have created a generic rule-based system for recognizing textual entailment that obtains competitive results on a real evaluation dataset. Our approach does not require training. We confront it with a strong baseline provided by the EDITS system (Kouylekov et al., 2011).

In Section 2 we describe the computational semantics framework that forms the basis of our approach. Section 3 details our entailment system, and in Section 4 we analyze our results from the task evaluation.

2 Minimal Recursion Semantics

Minimal Recursion Semantics (MRS) (Copestake et al., 2005) is a framework for computational semantics which provides expressive representations with a clear interface with syntax. MRS allows underspecification of scope, in order to capture the different readings of a sentence with a single MRS representation. We use the MRS analyses that are produced by the HPSG English Resource Grammar (ERG) (Flickinger, 2000).

The core of an MRS representation is a multiset of relations, called *elementary predications* (EPs). An EP represents a single lexeme, or general grammatical features. Each EP has a *predicate symbol*, and a *label* (also called *handle*) that identifies the EPs position within the MRS structure. Each EP contains a list of numbered arguments: ARG0, ARG1, etc., whose values are scopal or non-scopal variables. The ARG0 value is called the EP's *distinguished variable*, and denotes an event or state, or an entity.

Finally, an MRS has a set of *handle constraints* which describe how the scopal arguments of the EPs can be equated with EP labels. A constraint $h_i =_q h_j$ denotes equality modulo quantifier insertion. EPs are directly and indirectly linked through handle constraints and variable sharing, and the resulting MRS forms a connected graph.

In Figure 1, we see an MRS for the sentence *A woman is cutting a potato*. The topmost EP, _cut_v_1, has a list of three argument-value pairs: its distinguished variable e_3 denotes an event, and the variables x_6 and x_9 refer to the entities filling the agent and patient roles in the verb event. x_6 and x_9 are in turn the distinguished variables of the EPs that represent *a woman* and *a potato*, respectively.

3   System Description

In the following, T_{sent} and H_{sent} refer to the text and hypothesis sentence, and T_{mrs} and H_{mrs} to their MRS representations.

The core of our system is a rule based component, which bases entailment decision on graph alignment over MRS structures. An earlier version of the system is described in Lien (2014). The earlier version was developed on the data set from the SemEval-2010 shared task Parser Evaluation using Textual Entailment (PETE) (Yuret et al., 2010). Using no external linguistic resources, the system output positive entailment decisions for sentence pairs where core nodes of the H_{mrs} could be aligned to nodes in T_{mrs} according to a set of heuristic matching rules. The system we present in this paper extends the earlier version by adding support for contradiction recognition, and by using lexical relations from WordNet.

For our participation in the entailment recognition task, first, we did an analysis of the SICK trial data. In the ENTAILMENT pairs, H_{sent} is a paraphrase over the whole or part of the text sentence.

The changes from T_{sent} to H_{sent} can be syntactic (e.g., active-passive conversion), lexical (e.g., synonymy, hyponymy-hypernymy, multiword expressions replaced by single word), or T_{sent} contains some element that does not appear in H_{sent} (e.g., T_{sent} is a conjunction and H_{sent} one of its conjuncts, a modifier in T_{sent} is left out of H_{sent}). In the CONTRADICTION category, the sentences of a pair are also basically the same or paraphrases, and a negation or a pair of antonymous expressions create the contradiction. The NEUTRAL pairs often have a high degree of word overlap, but H_{sent} cannot be inferred from T_{sent}. Our system accounts for many of these characteristics.

The system bases its decision on the results of two procedures: a) an *event relation match* which searches for an alignment between the MRSs, and b) a *contradiction cue* check. After running these procedures, the system outputs

1. ENTAILMENT, if the event relation matching procedure found an alignment, and no contradiction cues were found,

2. CONTRADICTION, if contradiction cues were found,

3. NEUTRAL, if neither of the above conditions are met.

The *event relation matching procedure* extends the one developed in Lien (2014) to account for the greater lexical variation in the SICK data. The procedure selects all the EPs in T_{mrs} and H_{mrs} that have an *event variable* as their ARG0—we call them *event relations*. These event relations mainly represent verbs, verb conjunctions, adjectives, and prepositions. For each event relation H_{event} in the hypothesis the procedure tries to find a matching relation T_{event} among the text event relations. We say that H_{event} matches T_{event} if:

1. they represent the same lexeme with the same part-of-speech, or if both are verbs and H_{event} is a synonym or hypernym of T_{event}, and

2. all their arguments match. Two event relation arguments in the same argument position match if:

   - they are the same or synonymous, or the H_{event} argument is a hypernym of the T_{event} argument, or

$$\langle\, h_1,$$
$$h_4{:}\_\mathsf{a\_q}\langle 0{:}1\rangle(\text{ARG0 } x_6, \text{RSTR } h_7, \text{BODY } h_5),$$
$$h_8{:}\_\mathsf{woman\_n\_1}\langle 2{:}7\rangle(\text{ARG0 } x_6),$$
$$h_2{:}\_\mathsf{cut\_v\_1}\langle 11{:}18\rangle(\text{ARG0 } e_3, \text{ARG1 } x_6, \text{ARG2 } x_9),$$
$$h_{10}{:}\_\mathsf{a\_q}\langle 19{:}20\rangle(\text{ARG0 } x_9, \text{RSTR } h_{12}, \text{BODY } h_{11}),$$
$$h_{13}{:}\_\mathsf{potato\_n\_1}\langle 21{:}28\rangle(\text{ARG0 } x_9)$$
$$\{\, h_{12} =_q h_{13},\ h_7 =_q h_8,\ h_1 =_q h_2 \,\} \,\rangle$$

Figure 1: MRS for *A woman is cutting a potato* (pair 4661, SICK trial data).

- the argument in T_{event} represents a noun phrase and the argument in H_{event} is an underspecified pronoun like *somebody*, or
- the argument in T_{event} is either a scopal relation or a conjunction relation, and one of its arguments matches that of H_{event}, or
- the argument in H_{event} is not expressed (i.e., it matches the T_{event} argument by default)

The matching procedure does not search for more than one alignment between the event relations of H_{mrs} and T_{mrs}.

The *contradiction cue procedure* checks whether the MRS pairs contain relations expressing negation. The quantifier `_no_q_rel` negates an entity (e.g., *no man*), whereas `neg_rel` denotes sentence negation. If a negation relation appears in one but not the other MRS, we treat this as an indicator of CONTRADICTION.

Example: Figure 1 shows the MRS analysis of the hypothesis in the entailment pair *A woman is slicing a potato* $\Rightarrow$ *A woman is cutting a potato*. There is only one event relation in H_{mrs}: `_cut_v_1`. T_{mrs} is an equivalent structure with one event relation `_slice_v_1`. Using Word-Net, the system finds that `_cut_v_1` is a hypernym of `_slice_v_1`. Then, the system compares the `ARG1` and `ARG2` values of the event relations. The arguments match since they are the same relations. There are no contradiction cues in either of the MRSs, so the system correctly outputs ENTAILMENT.

If we look at the rule based component's output (Table 1) for the 481 of the 500 SICK trial sentence pairs for which the ERG produced MRSs, we get a picture of how well it covers the phenomena in the data set:

Of the 134 ENTAILMENT pairs, 59 were paraphrases where the variation was relatively limited

	gold ENT	gold CON	gold NEU
sys ENT	**59**	0	1
sys CON	0	**51**	14
sys NEU	75	22	**259**

Table 1: Output for the system on SICK trial data.

and could be captured by looking for synonyms, hyponyms, and treating the hypothesis as a subgraph of the text. The simple contradiction cue check, which looks for negation relations, covered 51 of 73 CONTRADICTION pairs.

75 ENTAILMENT and 22 CONTRADICTION pairs were not captured by the matching and contradiction cue procedures. Almost 30% of the ENTAILMENT pairs had word pairs whose lexical relationship was not recognized using Word-Net (e.g.: *playing a guitar* $\Rightarrow$ *strumming a guitar*). In the other pairs there were alternations between simple and more complex noun phrases (*protective gear* $\Rightarrow$ *gear used for protection*), change of part-of-speech from T_{sent} to H_{sent} for the same meaning entities (*It is raining on a walking man* $\Rightarrow$ *A man is walking in the rain*); some pairs required reasoning, and in some cases H_{sent} contained information not present in T_{sent}. In some cases, entailment recognition fails because the MRS analysis is not correct (e.g., misrepresentation of passive constructions).

The contradiction cue check did not look for antonymous words and expressions, and this accounts for almost half of the missing CONTRADICTION pairs. The rest contained negation, but were misclassified either because an incorrect MRS analysis was chosen by the parser or because synonymous words within the scope of the negation were not recognized.

EDITS We used a backoff-system for the pairs when the rule-based system fails to produce re-

System	1 Rules Only	2 Rules Only	3 Combined	4 Combined	5 Edits
Training	76.13	75.4	76.62	76.62	74.78
Test	77.0	76.35	77.12	77.14	74.79

Table 2: Submitted system accuracy on training and test set.

sults. Our choice was EDITS[1] as it provides a strong baseline system for recognizing textual entailment (Kouylekov et al., 2011). EDITS (Kouylekov and Negri, 2010) is an open source package which offers a modular, flexible, and adaptable working environment for experimenting with the RTE task over different datasets. The package allows to: *i)* create an entailment engine by defining its basic components; *ii)* train this entailment engine over an annotated RTE corpus to learn a model and *iii)* use the entailment engine and the model to assign an entailment judgment and a confidence score to each pair of an unannotated test corpus.

We used two strategies for combining the rule-based system with EDITS: Our first strategy was to let the rule-based system classify those sentence pairs for which the ERG could produce MRSs, and use EDITS for the pairs were we did not have MRSs (or processing failed due to errors in the MRSs) . The second strategy was to mix the output from both systems when they disagree. In this case we took the ENTAILMENT decisions from the rule-based, and EDITS contributes with CONTRADICTION and NEUTRAL.

4 Analysis

We have submitted the results obtained from five system configurations. The first four used the rule-based system as the core. The fifth was a system obtained by training EDITS on the training set. We use the fifth system as a strong baseline. In the few cases in which the rule-based system did not produce result (2% of the test set pairs) EDITS judgments were used in the submission. In *System 1* and *System 2* we have used the first combination strategy described in the end of section 3. In *System 4* and *System 5* the entailment decisions are a combination of the results from the rule-based system and EDITS as described in the second strategy in the same section. The rule-based component in *System 1* and *System 3* has more fine-grained

	Precision	Recall	F-Measure
Contradiction	0.8422	0.7264	0.78
Entailment	0.9719	0.4158	0.5825
Neutral	0.7241	0.9595	0.8254

Table 3: Performance of System 1.

negation rules so that _no_q_rel is not treated as a contradiction cue in different contexts (e.g., *No woman runs* does not contradict *A woman sings*). Table 2 shows the results for the five submitted systems.

The results demonstrate that the rule-based system can be used as a general system for recognizing textual entailment. It surpasses with 3 points of accuracy EDITS, which is an established strong baseline system. We are quite content with the results obtained as we did not use the training dataset to create the rules, but only the trial dataset. The combination of the two systems brings a slight improvement.

Overall the rule-based system is quite precise as demonstrated in Table 3. The numbers in the table correspond to *System 1* but are comparable to the other rule-based systems 2, 3 and 4. The system achieves an excellent precision on the entailment and contradiction relations. It is almost always correct when assigning the entailment relation. And it also obtains a decent recall, correctly assigning almost half of the entailment pairs. On the contradiction relation the system also obtained a decent result, capturing most of the negation cases.

5 Conclusions

Using a state-of-the-art semantic analysis component, we have created a generic rule-based system for recognizing textual entailment that obtains competitive results on a real evaluation dataset. An advantage of our approach is that it does not require training. The precision of the approach makes it an excellent candidate for a system that uses textual entailment as the core of an intelligent search engine.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A Survey of Paraphrasing and Textual Entailment Methods. *J. Artif. Intell. Res. (JAIR)*, 38:135–187.

Richard Bergmair. 2010. *Monte Carlo Semantics: Robust Inference and Logical Pattern Processing with Natural Language Text*. Ph.D. thesis, University of Cambridge.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal Recursion Semantics: An Introduction. *Research on Language & Computation*, 3(2):281–332.

Dan Flickinger. 2000. On Building a More Effcient Grammar by Exploiting Types. *Natural Language Engineering*, 6(1):15–28.

Milen Kouylekov and Matteo Negri. 2010. An Open-Source Package for Recognizing Textual Entailment. In *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010) ,Uppsala, Sweden*, pages 42–47.

Milen Kouylekov, Yashar Mehdad, and Matteo Negri. 2011. Is it Worth Submitting this Run? Assess your RTE System with a Good Sparring Partner. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment, Edinburgh Scotland*, pages 30–34.

Elisabeth Lien. 2014. Using Minimal Recursion Semantics for Entailment Recognition. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 76–84, Gothenburg, Sweden, April.

Andreas Wotzlaw and Ravi Coote. 2013. A Logic-based Approach for Recognizing Textual Entailment Supported by Ontological Background Knowledge. *CoRR*, abs/1310.4938.

Deniz Yuret, Aydin Han, and Zehra Turgut. 2010. SemEval-2010 Task 12: Parser Evaluation using Textual Entailments. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 51–56.