UAIC Participation at RTE-7

Mihai-Alex Moruz

"Al. I. Cuza" University, Faculty of Computer Science, Iasi, Romania

mmoruz@info.uaic.ro

Abstract

This paper describes the fifth participation of the UAIC textual entailment engine at the RTE shared task. Our approach is rule based and makes use of the notion of predicational semantics for detecting entailment. The system used is based on the system built for the RTE-6 challenge, which we have further modified and improved for the current task. The system works by attempting to match every entity in the hypothesis to at least one entity in the text, but for this version of the systems, the matching process is predication driven, which is to say we used predicates in T and H as pivot points for determining matching entities. Matching is achieved by using extensive semantic knowledge from suck knowledge bases as DIRT, VerbNet, WordNet, VerbOcean, Wikipedia and the Acronym database.

1. Introduction

One of the most relevant phenomena in natural language is that of variability, which can be loosely defined as stating the same ideas in different ways. While natural language variability can be addressed locally by each application, a more general solution is the notion of textual entailment, which was first introduced by (Dagan and Glickman, 2004):"textual entailment (entailment, in short) is defined as a relationship between a coherent text T and a language expression, which is considered as a hypothesis, H. We say that T entails H (H is a consequent of T), denoted by T⇒H, if the meaning of H, as interpreted in the context of T, can be inferred from the meaning of T."

The Recognizing Textual Entailment (RTE) task consists of creating a system that, given two pieces of text, can determine if the meaning of one text is entailed, or can be deduced from the other text. Although the basic definition for entailment remains the same, successive RTE challenges have grown more and more complex, as they have adapted to the requirements of such NLP applications as machine summarization and knowledge base population. For the RTE-7 challenge, the main task remained relatively unchanged from that of RTE-6, namely "Given a corpus, a hypothesis H, and a set of "candidate" sentences retrieved by Lucene from that corpus for H, RTE systems are

required to identify all the sentences that entail H among the candidate sentences". This paper describes the system we have used in the RTE-7 challenge.

The rest of the paper is structured as follows: chapter 2 describes the system we have used for the RTE-7 challenge, chapter 3 discusses results, and chapter 4 gives a set of conclusions.

2. System description

In the course of analysis of various RTE datasets and examples, we have come up with the intuition that entailment pairs can be solved, in the majority of cases, by examining two types of information, which lead to a semantic understanding of the text and the hypothesis (Moruz, 2011):

- The relation of the predicates in the hypothesis to the ones in the text. In this context, predicates are taken to mean the verb itself, together with its arguments and adjuncts; thus, the comparison of two predicates is a comparison of complex structures, which are, in essence, atomic propositions (clauses). If the verbs, together with their arguments and adjuncts, match over T and H, we have ENTAILMENT (by matching we understand that $\forall$ verb q $\in$H,$\exists$ verb p$\in$T so that p$\rightarrow$q; we say that a predicate p entails a predicate q, p$\rightarrow$q, if q is a consequence of p, or p and q are synonyms, or q is a subevent of p).
- Each argument or adjunct is an entity, with a set of defined properties. It may happen that, despite agreement at the level of the verb, there may be differences in terms of entity properties for similar arguments or adjuncts. In order to solve such cases correctly, each argument and adjunct is considered an entity with an attached set of attributes, and only if they match we have ENTAILMENT. By matching at the argument level we understand that, given the feature sets for the arguments arg_T and arg_H, the unification of these feature structures (as defined in unification grammars) is successful and is equal to the feature structure of arg_T.

This operational definition for textual entailment is based on the notion that the generator of each sentence is a predicate; since all the information of an entailed hypothesis must be contained in the text, it follows that all the predicates in the hypothesis must subsume some predicate in the text. Another advantage of this interpretation for the definition of textual entailment is the fact that it can

be applied directly to an existing rule based system, in the sense that the matching rules need to be applied so that they attempt to determine the subsumption relation required for proving entailment.

The basis of our implementation is the system described in (Iftene, 2008), (Iftene and Moruz, 2009), which was further developed and modified according to the interpretation of the entailment definition given above. The general architecture of the system is similar to that of our previous systems, and is given in Fig. 1.

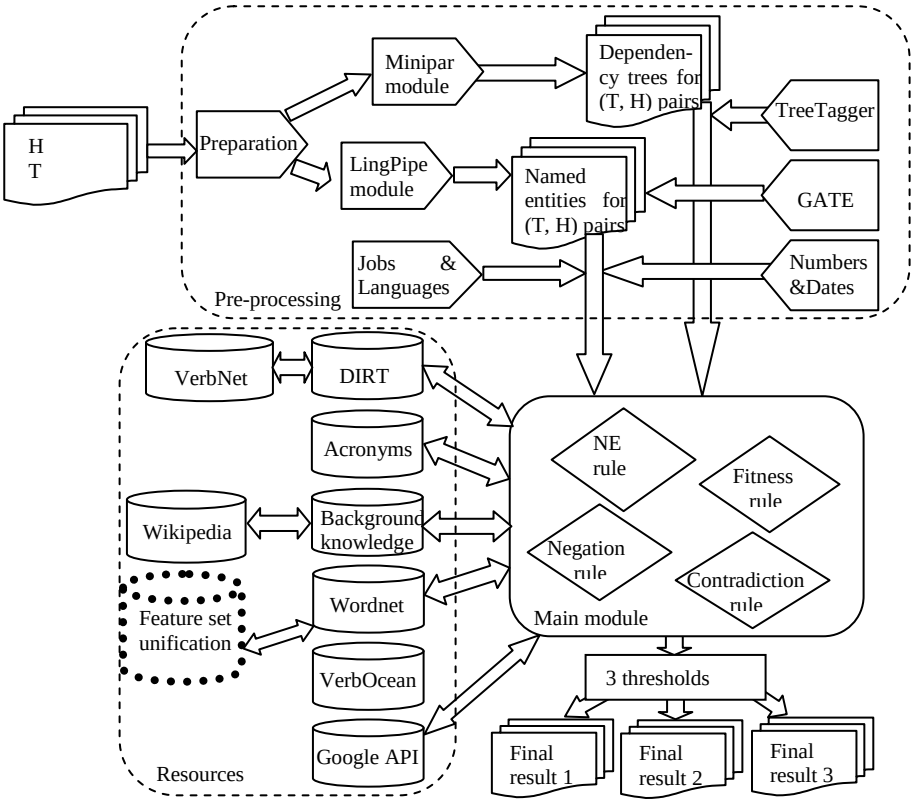

Figure 1: RTE-7 System architecture

The dotted module is the newly added feature set unification module, which is responsible for argument alignment. Verb alignment is performed using VerbNet, and, in those cases where VerbNet did not provide coverage, DIRT. The pre-processing and the rule modules are largely the same as those described in (Iftene and Moruz, 2009, 2010). The pre-processing step includes a syntactic parser (MINIPAR – (Lin, 1998)), a named entity recognizer (LingPipe[1] coupled with GATE (Cunningham et al., 2001)) and a series of lexical transformations that are applied to the input texts. The

[1] LingPipe: http://www.alias-i.com/lingpipe/

main module is responsible for applying the entailment rules that we have defined; it first attempts to find entailing matches for verbs, and then attempts to find entailing matches for the arguments of the verbs. The quality of the matches is translated into a local fitness score, which is then used for computing a global fitness score for the T-H pair. Empirically determined thresholds are then used to classify the entailment result by means of global fitness. The resource module includes BK extracted from Wikipedia, verb oriented resources such as VerbNet (Kipper-Schuler, 2005), DIRT (Lin and Pantel, 2001) and VerbOcean (Chklovski and Pantel, 2004), an acronym database and WordNet (Fellbaum, 1998).

3. Results in RTE-7

Using the system described in section 2, we participated in the main task of the RTE-7 evaluation campaign with three distinct runs, obtained by running the system with different thresholds. The results are given in table 1 below:

Run ID	Precision	Recall	F-Measure
001	45.40%	18.12%	25.90%
002	30.21%	25.84%	27.85%
003	18.04%	29.66%	22.43%

Table 1: Results for the RTE-7 Main Task

The first run was obtained with the thresholds set to maximize precision at the expense of recall. Run two was obtained by lowering the threshold for separating the entailment and non-entailment cases; this is the reason for the higher recall and the lower precision. The third run was obtained by further lowering the value of the threshold and thus the justification for the low precision.

Even though the results we have obtained are an improvement over last year's submissions, they are still quite low. The main reason for the low performance is our low recall, which suggests that the system is excessively restrictive when selecting entailment candidates.

The relevance of each of the components was also tested by means of ablation tests, which were required by the organizers. The results of the ablation tests for the second submitted run (our best scoring run) are given in table 2 below.

System Description	RTE-7				
	P (%)	R (%)	F (%)	C (%)	WR (%)
Without verb resources	29.78	26.45	28.02	-0.17	-0.61

System Description	RTE-7				
	P (%)	R (%)	F (%)	C (%)	WR (%)
Without BK	30.21	25.84	27.85	0	0
Without NE resources	28.60	49.08	36.14	-8.29	-29.76
Without the Negation rule	30.09	26.91	28.41	-0.56	-2.01
Without the Contradiction rule	30.21	25.84	27.85	0	0

Table 2: Components' relevance for RTE-7 main task

The meanings of the columns are the following:

- $Precision_{Without\_Component}$;
- $Recall_{Without\_Component}$;
- $F\text{-}measure_{Without\_Component}$;
- $Contribution_{Component} = Full\_system\_F\text{-}measure - F\text{-}measure_{Without\_Component}$
- $WeightedRelevance_{Component} = \frac{100 \times Contribution_{Component}}{Full\_system\_f-measure}$

As can be seen in Table 2, most of the components actually decrease the performance of the system for Run 2, but the largest decrease is brought by the NE component. This is due to the fact that the NE rules are far too restrictive, in the sense that only exact matches over names are allowed (partial matches are penalized, and the absence of matches invariably means no entailment). In order to solve this, we need to perform some form of nominal coreference resolution and to expand our acronym database.

The rest of the ablated resources either brought no change at all or slightly decreased performance. We have not yet determined the exact reason for their poor performance, but preliminary analysis suggests that the very restrictive application of the rules is responsible. This conclusion is mainly supported by the comparatively high precision for the first two runs, and the rather recall for all submitted runs.

4. Conclusions

Even though the results we have obtained for the RTE-7 challenge are an improvement over those obtained in RTE-6, the excessive restrictiveness with which we have applied our entailment rules prevented the system from fully using the resources available, which greatly reduced its performance. For the future versions of the system we intend to allow for more flexibility in determin-

ing matching entities, and to further extend NE and acronym resources, in order to improve the results of the application of the NE rule.

Acknowledgments

The author of this paper thanks the members of the NLP group in Iasi for their help and support at different stages of the system development.

References

Chklovski, T., Pantel, P. 2004. *VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations*. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04). Barcelona, Spain.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. 2001. *GATE: an architecture for development of robust HLT applications.* In ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2001, 168--175, Association for Computational Linguistics, Morristown, NJ, USA.

Fellbaum, C. 1998. *Wordnet: An electronic lexical database*. MIT Press, Cambridge, Mass.

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

Iftene, A. 2008. *UAIC Participation at RTE4*. In Text Analysis Conference (TAC 2008) Workshop - RTE-4 Track. National Institute of Standards and Technology (NIST). November 17-19, 2008. Gaithersburg, Maryland, USA.

Iftene, A., Moruz M. A. 2009. *UAIC Participation at RTE-5*, In Text Analysis Conference (TAC 2009) Workshop - RTE-5 Track. National Institute of Standards and Technology (NIST). November 16-17, 2009. Gaithersburg, Maryland, USA.

Iftene, A., Moruz M. A. 2010. *UAIC Participation at RTE-6*, In Text Analysis Conference (TAC 2010) Workshop - RTE-6 Track. National Institute of Standards and Technology (NIST). November 15-16, 2010. Gaithersburg, Maryland, USA.

Kipper-Schuler, K. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia, PA, June

Lin, D. 1998. *Dependency-based Evaluation of MINIPAR*. In Workshop on the Evaluation of Parsing Systems, Granada, Spain, May, 1998.

Lin, D., Pantel, P. 2001. *DIRT - Discovery of Inference Rules from Text*. In Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-01). pp. 323-328. San Francisco, CA.

Moruz, M. A. 2011. *Predication Driven Textual Entailment*, PhD Thesis, Faculty of Computer Science, "Al. I. Cuza" University, Iasi, Romania