FBK Participation in the RTE-7 Main Task

Yashar Mehdad[1,2], Matteo Negri[1], José Guilherme Camargo de Souza[1,2]
Alina Petrova[3]
[1]FBK-irst, [2]University of Trento
[3]Technical University of Dresden
{mehdad,negri,desouza}@fbk.eu, alina.v.petrova@gmail.com

Abstract

This paper overviews FBK's participation in the RTE-7 *Main* task organized within the Text Analysis Conference (TAC) 2011. Our participation is characterized by two main themes, namely:

1. The attempt to move from token-level overlap measures (*i.e.* a count of the terms in the *hypothesis* that can be mapped to terms in the *Text*), to phrase-level overlap measures that take into account a larger context to favour system's precision;
2. The attempt to use paraphrase tables derived from *parallel data* as the main source of lexical knowledge for the mapping.

In contrast with previous experiments over different datasets on one side, and the scores achieved over the RTE-7 DEV_SET on the other side, our final results are lower than those obtained with the simpler token-overlap algorithm (41.90% Vs 44.1% Micro-Averaged F-measure). The motivations for this unexpected performance drop are still under investigation.

1 Introduction

Building on the outcomes of previous participations to the RTE challenge [1, 2], our approach to the 2011 edition of the RTE task moves from the observation that word-overlap measures represent a strong, hard to beat baseline for recognizing textual entailment. Along this direction, our previous works considered the overlap between texts and hypotheses only at the level of *single words*, taking advantage of lexical resources and methods biased towards tokens [3, 4, 5]. Although quite effective, and always leading to above average results [6], token-overlap could be in principle easily improved by considering *phrases* (*i.e.* more context) to enhance precision. To overcome the bias towards single words, we recently explored the viability of *n-gram*-based solutions that exploit paraphrase tables extracted from parallel data [7]. The intuition behind the two approaches is the same but, instead of measuring the amount of words in H that can be mapped to words in T (either directly, or through entailment-preserving relations provided by lexical knowledge sources), our RTE-7 system measures the amount of n-grams in H that can be mapped to n-grams in T.

In addition to this change, we explored the integration within a Machine Learning framework of a number of features that reflect a more linguistically motivated representation of the T-H pairs. To this aim, information about the semantic similarity between texts and hypotheses, semantic roles, named entities, and semantic relations between entities has been considered during system's development.

The final submitted runs have been produced by: *i)* running n-fold cross validation on the RTE-7 DEV_SET with different algorithms and sets of features in order to select the most promising/stable configuration of our system, and *ii)* using, as a term of comparison, the score obtained by the EDITS system [6, 8] on the same dataset. In both cases, our pre-submission results were very positive, showing: *i)* significant improvements over the simpler token-overlap, *ii)* the positive effect of calculating phrasal-overlap by means of paraphrase tables derived from parallel data, and *iii)* the usefulness, beyond word overlap counts, of other semantically motivated features. In spite of this, the official scores achieved over the TEST_SET are significantly lower than expected. First, the performance drop on test data is larger than the decrease observed when token-overlap is used as a single feature. Second, the results of the *post hoc* ablation tests contradict some of the findings we made during training (*e.g.* about the usefulness of some of the features). Although error analysis is still underway, both conclusions confirm the difficulty, in the RTE task, of improving over relatively simple baselines even with promising, linguistically motivated approaches.

The following sections overview our participation in the RTE-7 Main task, providing details about the experiments carried out at the training stage, the submitted runs, the results achieved, and the *post-hoc* ablation tests performed to check the validity of what we learned during training.

2 Training, submissions, and results

The RTE-7 Main task has the same formulation as in RTE-6, that is: "Given a corpus *C*, a hypothesis *H*, and a set of "candidate" entailing sentences for that *H* retrieved from *C* by the Lucene search engine, the RTE-7 main task consists in identifying all the sentences that entail *H* among the candidate sentences".

2.1 Training the system

System's training was carried out through the following steps:

RTE corpus creation. As a first step we created a set of entailment pairs of the type $TCand_x$-*H* for each hypothesis *H* and for each candidate sentence for that *H*.

Filtering. The resulting entailment corpus, highly unbalanced towards negative examples (*i.e.* 20284 "NO" entailment pairs, Vs 1136 positive examples) was then processed in order to reduce such disproportion. To this aim, by filtering out pairs composed by candidate entailing sentences with a low Lucene score ($\leq$ 0.02), the original corpus was reduced to 1101 positive and 13614 negative examples. The remaining pairs were further filtered by considering the lexical overlap between texts and hypotheses. Such

overlap was calculated applying a modified version of the Lesk measure which sums, for each n-gram level (from 1 to 5), the squares of the length of the phrasal matches.

$$Lesk_{mod} = \frac{\sum_{n=1}^{5} \#matched(n) * n^2}{length(H)^2}$$

Considering the asymmetry of the entailment relation, the score is normalized by the square of the length of H, instead of the product of the lengths of the strings as it is done by the original Lesk formula. With a threshold manually set to 0.015, the modified Lesk measure allowed to obtain a final RTE corpus composed of 999 positive and 9086 negative pairs (respectively 137 and 11198 less than the original ones).

Preprocessing. The entailment corpus resulting from the previous step was processed using the TreeTagger [11] to obtain tokenization, lemmatization and part-of-speech tagging. In order to investigate also higher level linguistically motivated features, we further processed the created corpus at the level of dependency tree representations [14], named entities [13], semantic roles [12] and dependency relation triples [15]. For each T-H pair, these different levels of information were used to extract the features discussed in the next section.

Feature extraction and selection. In our approach, the entailment recognition process builds on feature vector representations of the T-H pairs, which are used to learn a reliable model for the classification of new unseen pairs. These features try to capture various phenomena affecting the entailment decision. The features extracted can be grouped as follows:

- **Lexical overlap**, including Word-Overlap (WO), Longest Common Sub-sequence (LCS) [8] and Rouge (RG) [16]: these features capture the presence of H terms (tokens or lemmas) in T.
- **Phrasal matching (PPT)** using paraphrase tables [7]: these features capture the presence of H n-gram (n = 1 .. 5) phrases (tokens or lemmas or stems) in T on the basis of a paraphrase table extracted from bilingual corpora.
- **Semantic role matching (SR)** [12, 18]: these features capture similarities between T and H by comparing the semantic roles (*i.e.*, arguments and adjuncts) which occur in the predicates.
- **Named entity matching (NE)** [13, 18]: these features analyze similarities between T and H by comparing the named entities, according to their type, which co-occur in them.
- **Dependency relation matching (RELEX)** [15]: these features capture similarities between T and H by comparing dependency relation triples.
- **Wikipedia similarity score (WIKI)** [3, 17]: this feature estimates the similarity of two strings (T and H) using a latent semantic model which is built over Wikipedia.

In order to select the most effective feature set, among the wide range of algorithms for learning from unbalanced datasets we used the logistics and optimized decision tree algorithms available in WEKA [19]. Moreover, we set the algorithms to maximize the F-measure over positive entailment pairs as a learning criterion.

We tested the algorithms mentioned above over different sets of features in various combinations. In order to select the best features for the final runs, we performed learning and classification using 10-fold and 2-fold cross validation. The results achieved over the DEV_SET, under both n-fold validation conditions, showed the positive contribution of most of the features we experimented with. Apart from *semantic roles* (SR) and *dependency relations* (RELEX), which we are not able to fully exploit yet, all the other features contributed to increase baseline performance (up to 3.6 F-measure points with 10-fold cross validation). The scores achieved with the best performing configurations (*i.e.* those used for the final submission, described in the following section) are reported in Table 1. The results calculated over the same dataset with the latest release of the EDITS system (EDITS-GA), used as a term of comparison, are also reported in the table.

2.2 Submitted runs and results

Building on the results obtained over the DEV_SET, for the RTE7 Main task we submitted the following three runs:

Run 1.
Features: WO, LCS, PPT, WIKI, NE.
Algorithm: logistics optimized on F-measure of positive entailment pairs.
Training dataset: filtered using Lucene and Lesk score.

Run 2.
Features: WO, LCS, PPT, WIKI, NE.
Algorithm: decision tree applying the LogitBoost strategy optimized on F-measure of positive entailment pairs.
Training dataset: filtered using Lucene and Lesk score.

Run 3.
Features: WO, LCS, PPT, WIKI, NE.
Algorithm: logistics optimized on F-measure of positive entailment pairs.
Training dataset: unfiltered (*i.e.* all the original DEV_SET).

The results achieved by each run, both on the training and test data, are reported in Table 1. As can be seen from the table, our best result has been achieved by Run 3 (41.90% Micro-Averaged F-measure). It has to be observed that, during training, such configuration achieved significantly lower results compared to the others. Moreover, compared to *i)* EDITS, *ii)* 10-fold, and *iii)* 2-fold cross validation scores obtained on the DEV_SET, all these scores reveal a considerable performance drop, larger than the (somehow) expected effect of data overfitting.

	10-fold (Dev)	2-fold (Dev)	EDITS-GA (Dev)	Test (official)
Run 1 (filtered)	52.8	56.1	52.2	39.89
Run 2 (filtered)	53.1	50.7	52.2	39.50
Run 3 (unfiltered)	51.2	50.8	47.6	41.9

Table 1: F-measure scores over Dev (10-fold & 2-fold cross validation) compared with EDITS-GA, and official Test results.

3 Ablation tests

As in previous editions of the RTE Main Task, also this year ablation tests were required for systems participating in the Main task. In order to better understand the impact of different knowledge resources and tools, participants were required to run their system over the test data removing one resource at a time. To this aim, we submitted four additional runs. Each run has been obtained applying our best model trained on the unfiltered DEV_SET (the best set according the official results), removing the most useful features according to the training results[1]. Table 2 shows the results achieved by the four ablation tests, reporting the ablated resource(s), and the 10-fold validation scores achieved, with the same configuration, over the unfiltered DEV_SET.

Ablation test No.	Ablated run	Ablated resource	10-fold (Dev)	Test
-	Run 3	-	51.2	41.9
1	Run 3	Paraphrase Table	50.2	43.33
2	Run 3	Named Entities feature	49.2	42.79
3	Run 3	Wiki similarity feature	49.4	44.5
4[1]	Run 3	All but token-overlap	47.6	44.1

Table 2: Ablation results

As can be seen from the table, the ablation tests results contradict the observations made with n-fold cross validation over the training data. While on the DEV_SET each resource contributes to a significant improvement over the baseline configuration (using only token-overlap), on the TEST_SET they all have a negative impact. Among others, these results contradict previous findings about the effectiveness of: *i)* computing overlap measures at the *phrasal* level [7], and *ii)* computing similarity between T and H using LSA models built over Wikipedia [17].

4 Conclusion

Our approach to the 2011 edition of the RTE task moved from, and ended with the observation that word-overlap measures represent a strong, hard to beat baseline for recognizing textual entailment. Despite the attempt to improve over simple token-overlap computation with a more linguistically motivated approach (also based on a

[1] Although the 4^{th} run is not compliant with the specifications for ablation tests (all resources are indeed ablated), we submitted it to eventually check the performance of a basic system based on token-overlap.

richer representation of the T-H pairs), the final evaluation results led to contradictory conclusions that confirm the difficulty of the task.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n. 248531 (CoSyne project).

References

[1] Yashar Mehdad, Matteo Negri, Elena Cabrio, Milen Kouylekov, and Bernardo Magnini (2009). Using Lexical Resources in a Distance-Based Approach to RTE In *TAC 2009 Notebook Papers* Gaithersburg, MD, US.

[2] Milen Kouylekov, Yashar Mehdad, Matteo Negri, and Elena Cabrio. (2009). FBK Participation in RTE6: Main and KBP Validation Task. In *Proceedings of the Sixth Recognizing Textual Entailment Challenge.* Gaithersburg, MD, US.

[3] Milen Kouylekov, Yashar Mehdad, and Matteo Negri (2010). Mining Wikipedia for Large-Scale Repositories of Context-Sensitive Entailment Rules. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2010).*

[4] Christiane Fellbaum, editor (1998). WordNet: An Electronic Lexical Database. Language, Speech and Communication. MIT Press.

[5] Chklovski, T. and Pantel, P. (2005). 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.

[6] Milen Kouylekov, Matteo Negri, and Yashar Mehdad (2010). 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.

[7] Yashar Mehdad, Matteo Negri, and Marcello Federico (2011) Using Bilingual Parallel Corpora for Cross-Lingual Textual Entailment. *Proceedings of ACL-HLT 2011*.

[8] Milen Kouylekov, and Matteo Negri (2010). An Open-Source Package for Recognizing Textual Entailment. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010) System Demonstrations* Uppsala, Sweden, July 11-16.

[9] Yashar Mehdad, Matteo Negri, and Marcello Federico (2010) Towards Cross-Lingual Textual Entailment. *Proceedings of NAACL-HLT 2010*.

[10] D. Klein and C.D. Manning (2003) Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems 15* Cambridge, MA. MIT Press.

[11] Helmut Schmid Probabilistic Part-of-Speech Tagging Using Decision Trees 1994.

[12] Surdeanu, M., & Turmo, J. (2005) Semantic Role Labeling Using Complete Syntactic Analysis. In *Proceedings of CoNLL Shared Task 2005.*

[13] Surdeanu, M., Turmo, J., & Comelles, E. (2005) Named Entity Recognition from Spontaneous Open-Domain Speech. In *Proceedings of the 9th Interspeech 2005.*

[14] Lin, D. (1998) Dependency-based Evaluation of MINIPAR. In *Proceedings of the Workshop on the Evaluation of Parsing Systems.*

[15] Mike Ross, Linas Vepstas, and Ben Goertzel (2005) Relex semantic relationship extractor. *http://opencog.org/wiki/RelEx*

[16] Lin, Chin-Yew 2004. ROUGE: A Package For Automatic Evaluation Of Summaries. In *Workshop On Text Summarization Branches Out 2004.*

[17] Yashar Mehdad, Alessandro Moschitti and Fabio Massimo Zanzotto. Syntactic/Semantic Structures for Textual Entailment Recognition. In *Proceedings of the HLT-NAACL, 2010, Los Angeles, Calfornia.*

[18] Jesús Giménez and Lluís Màrquez. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. The Prague Bulletin of Mathematical Linguistics, No. 94, 2010.

[19] Ian H. Witten and Eibe Frank and Len Trigg and Mark Hall and Geoffrey Holmes and Sally Jo Cunningham. Weka: Practical Machine Learning Tools and Techniques with Java Implementations. 1999.