A Semantic Approach To Textual Entailment: System Evaluation and Task Analysis

Aljoscha Burchardt, Nils Reiter, Stefan Thater
Dept. of Computational Linguistics
Saarland University
Saarbrücken, Germany
{albu,reiter,stth}@coli.uni-sb.de

Anette Frank*
Dept. of Computational Linguistics
University of Heidelberg
Germany
frank@cl.uni-heidelberg.de

Abstract

This paper discusses our contribution to the third RTE Challenge – the SALSA RTE system. It builds on an earlier system based on a relatively deep linguistic analysis, which we complement with a shallow component based on word overlap. We evaluate their (combined) performance on various data sets. However, earlier observations that the combination of features improves the overall accuracy could be replicated only partly.

1 Introduction

This paper reports on the system we used in the third PASCAL challenge on Recognizing Textual Entailment. The system is based to a large extent on Burchardt and Frank's system (2006) used in the second RTE challenge (Bar-Haim et al., 2006); it relies on a relatively deep linguistic analysis, which we complement with a shallow component based on word overlap. As the system has been described earlier, we concentrate on a more systematic discussion of the system behaviour, aiming at spotting promising anchors for future extensions and improvements.

It has been observed for related systems that a combination of separately trained features in the machine learning component can lead to an overall improvement in system performance, in particular if features from a more "informed" component and shallow ones are combined (Hickl et al., 2006; Bos and Markert, 2006). We provide a detailed analysis of our system's behaviour on different training and test sets. However, we could not replicate the effects observed by others on all corpora – often, the accuracy of the combined features is not higher than the best individual features or feature sets. For the RTE 3 test set, the combined features actually lead to a slightly lower accuracy.

One candidate future enhancement of our system is to refine the relatively unrestricted graph matching that compares the analyses of text and hypothesis and underlies the definition of the deep features. But a more controlled, "rule based" definition of an adequate graph matching seems to rely on a deeper understanding of the notion of textual entailment.

In Section 2, we review the basic architecture of our system, and report on improvements and extensions. In Section 3, we provide a detailed evaluation of the system on different data sets. In Section 4, we report on some findings we made in a small annotation experiment we conducted at our department. In Section 5, we conclude and give a short outlook.

2 The SALSA RTE System

In this Section, we review the basic architecture of the SALSA RTE system, and report on some improvements and extensions. More details can be found in (Burchardt and Frank, 2006).

2.1 Architecture

The SALSA RTE system is based on three main components: (i) a linguistic analysis of text and hypothesis based primarily on LFG and Frame Semantics (Baker et al., 1998), (ii) the computation of a *match graph* that encodes the "semantic overlap" between text and hypothesis, and (iii) a statistical entailment decision.

*By the time of writing, Anette Frank was affiliated at Saarland University and DFKI Saarbrücken.

Proceedings of the Workshop on Textual Entailment and Paraphrasing, pages 10–15,
Prague, June 2007.

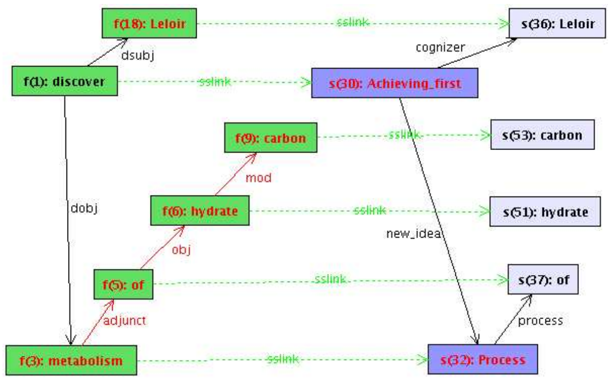

Figure 1: Linguistic Analysis of *Leloir discovered the metabolism of carbon hydrates.* (RTE3-test).

Linguistic analysis. The primary linguistic analysis components are the probabilistic LFG grammar for English developed at PARC (Riezler et al., 2002), and a combination of systems for frame semantic annotation: the probabilistic *Shalmaneser* system for frame and role annotation (Erk and Pado, 2006), and the rule-based *Detour* system for frame assignment (Burchardt et al., 2005).

Frame semantic analysis is especially interesting for the task of recognising textual entailment as it offers a robust yet relatively precise measure for semantic overlap. The lexical meaning of predicates and their arguments are modelled in terms of *frames* and *roles*. A frame describes a prototypical situation and roles identify participants involved in the situation. Frames provide normalisations over divers surface realisations, including variations in argument structure realisations. For instance, *buy*, *sell*, and *purchase* are all associated with the same frame.

The linguistic analysis combines LFG f-structures and FrameNet frames computed by the Shalmaneser and Detour systems, resulting in a projection from f-structures to a semantic layer of frames and "pseudo predicates" for f-structure predicates that do not project frames. Figure 1 shows the most important parts of the analysis of hypothesis 109 from the RTE 3 test set as an example. The LFG f-structure is shown on the left, and the dotted lines indicate the projection to semantic nodes on the right. The predicate *discover* is associated with the frame ACHIEVING_FIRST, the semantic role COGNIZER points to the pseudo-predicate *Leloir* and NEW_IDEA points to the PROCESS frame evoked by *metabolism*.

Semantic nodes are further projected into an ontological analysis layer containing WordNet (Fellbaum, 1997) senses and SUMO (Niles and Pease, 2001) classes. Semantic phenomena not treated by FrameNet like anaphora, negation or modality are (approximately) encoded with special operators. The resulting, layered graph structures for text and hypothesis thus provide access to the different types of information in a principled way.

Semantic overlap and match graphs. The system approximates textual entailment in terms of the "semantic overlap" between text and hypothesis. It compares their LFG f-structures with semantic and ontological projection by determining compatible, *matching* nodes and edges. The result is stored in a *match graph*, which contains all (pairs of) matched nodes and edges. Nodes can match if they are labelled with identical frames or predicates, or if the nodes are semantically related on the basis of WordNet or FrameNet frame relations. One node from the hypothesis may match multiple nodes from the text and vice versa. The matching of edges is restricted to edges that connect matching nodes, or nodes taking identical atomic values.

The match graph primarily encodes the "similarity" of text and hypothesis. In order to capture also a certain degree of "dissimilarity," nodes are deleted from the match graph if they occur in incompatible modality contexts.

Statistical entailment decision. Given a match graph and the graphs for text and hypothesis, we extract various features to train a machine learning model for textual entailment. In total, we extract 47 distinct features, which can be grouped according to their (i) level of representation (lexical, syntactic, semantic), (ii) degree of connectedness in the match graph, (iii) source (text, hypothesis or match graph), and (iv) proportional relation (hypothesis/text, match/hypothesis ratio).

2.2 Improvements

Sentence splitter. To cope with longer texts, we integrated the sentence splitter of the JTok tokeniser (Schäfer, 2005) into the system.

WordNet interface. The WordNet interface now treats particle verbs like *throw out* correctly. Moreover, we tested the usability of WordNet's verb entailment information as well as antonymy on nouns, verbs, and adjectives as basis for heuristic inferences in the graph matching process. However, for the given data, the number of text-hypothesis pairs where the relations are instantiated at all is marginal.

Frame semantic projection. The interface between the LFG and the Detour and Shalmaneser systems has been improved: by now 97% (+7%) of the frames and 74% (+10%) of the roles can be projected (on the RTE3 test set), resulting in an average of 6.6 frames and 5.5 roles per sentence.

2.3 Extensions

In addition to the above improvements, the system has been extended in two respects. We added a shallow component based on lexical overlap and a "meta learner" to study the combinatorics of machine learners' results.

Lexical overlap. In order to evaluate the performance of our system, we implemented a simple baseline system that approximates textual entailment in terms of lexical overlap between text and hypothesis. This shallow system is also used as a component to complement our full system in one of the two runs submitted to the RTE3 challenge.

The shallow system measures the relative number of words in the hypothesis that also occur in the text. Both text and hypothesis are tagged and lemmatised using Tree Tagger (Schmid, 1994), taking only nouns, non-auxiliary verbs, adjectives and adverbs into account. Training a decision tree on the relative word-overlap as single feature yields a system which performs comparable to earlier word-overlap based systems, achieving an accuracy of 60.6 % if trained and tested on the RTE 2 development and test set, respectively (using Weka's J48 classifier), or 57.5 % if we use Weka's LogitBoost classifier.

Weka Interface. Finally, we improved the machine learning back-end which feeds our extracted features into the Weka toolkit (Witten and Frank, 2005). This allows to train features in arbitrary combinations, with different machine learners.[1] Moreover, it supports testing the effect of using voting or a "meta learner" after training individual features or feature groups separately.

[1] The figures we present in the following are all computed with the LogitBoost classifier.

		IE	IR	QA	SUM
Run 1 (III)	62.25%	51%	68%	74%	57%
Run 2 (II)	62.62%	50%	69%	72.5%	59%

Table 1: Results of the SALSA RTE system (combined training set: RTE2-dev/-test and RTE3-dev).

3 Results

3.1 RTE3 Results

In the RTE3 task, we submitted two runs, one with (run 1) and one without (run 2) the lexical overlap component. Both achieved almost the same results, as can be seen in Table 1. The feature combinations (II, III) are explained below in detail.

3.2 Feature Combination

We investigated the behaviour of our systems on various combinations of three different sets of 800 text hypothesis pairs (RTE2-dev/-test and RTE3-dev). We tested four different feature configurations:

(I) All 47 features generated in our system (excluding the lexical overlap component)

(II) Three selected features of our system (run 2):

- Overlap of LFG predicates
- Matching of grammatical functions (deep subject/object, modifier, ...)
- Average size of connected parts ("clusters") of the match graph, comprising syntactic and semantic information

(III) The features from II plus lexical overlap (run 1)

(IV) Lexical overlap alone

In every configuration (except IV), the features were trained separately first, then a "meta classifier" was used to make the final entailment decision. We will use `Featureset-Training_set-Test_set` as notation for the configurations, e.g., `I-D2T2-D3` means all 47 features trained on the development and test set of RTE2, tested on the development set of RTE3. The results are shown in Table 2.

test → ↓ train	D2				T2				D3			
	I	II	III	IV	I	II	III	IV	I	II	III	IV
D2					56.25	57.25	58.625	57.5	57.875	61.125	66.375	66.625
T2	56.375	58.75	60.625	61.625					57.5	60.875	63.75	64.625
D3	53.875	61.25	61.75	61.75	56.625	58.75	57.25	57.25				
D2T2									58.5	64.25	65.875	66.375
D2D3					58	58.625	60	58.5				
T2D3	56.75	61.25	60.875	60.875								

Table 2: Performance of the different feature combinations on different training and test sets.

3.3 Corpus Variance

A general observation is, that almost all features behave quite differently on different training and test sets, as do feature combinations. Testing on T2 (RTE2-test) seems to be the hardest task. Not a single feature combination achieved an accuracy of more than 60%. In contrast, the best performance was 66.625% accuracy (IV-D2-D3). [2]

Usually, using a larger training set (the bottom part of table 2) should lead to a better performance.[3] However, this effect could not be observed here for all configurations. For most feature sets the performance gain is very small, e.g. from 56.375 (I-T2-D2) and 53.875 (I-D3-D2) to 56.75 (I-T2D3-D2). On some feature sets, the performance even drops, e.g. from 61.625 (IV-T2-D2) and 61.75 (IV-D3-D2) to 60.875 (IV-T2D3-D2). The largest boost occurred for feature set II. It's performance increased from 61.125 (II-D2-D3) and 60.875 (II-T2-D3) to 64.25 (II-D2T2-D3). It would be very interesting to see how the performance would develop on a much larger training set.

3.4 Feature Variance

The variance among individual features and feature sets is also large. Feature set II contains the most reliable and stable features. We tested how this "more informed" feature set (II) compares to the shallow word overlap feature (IV) and whether their combination (III) increases accuracy.

As can be seen from Table 2, in most of the cases, IV performs best, e.g. in the D2-D3 configuration, where feature set II alone achieves 61.125, while the combination with IV boosts the performance by 5% (III). On the other hand, there are cases, where the inclusion of the word overlap feature lowers the performance, e.g. from 61.25 (II-T2D3-D2) to 60.875 (III-T2D3-D2).

It is also interesting that combinations of features often perform lower than the best individual feature in the set. For instance, in D2T2-D3 III achieves 65.875, compared to 66.375 for IV alone. We generally could not observe a positive effect for the combination of features in a meta feature. In almost all configurations, the meta feature performed worse or equally well as the best individual feature. Apart from the size of the training data, feature dependence might be an explanation for this.

3.5 Task Variance

Figure 2 shows a per task analysis for the feature sets II, III and IV (D2T2-D3). The system performs best on the Question Answering task, where it achieves almost 80% accuracy. This differs from last year's experience, where the system performed best in the Summarization task. Given the general variance discussed above, this observation does not seem to allow general conclusions.

Again, there is a large variability of the overlap feature as well, which ranges between 52.5% (IE) and 79% (QA). This variability can partly be explained if we compare the average word overlap measures for positive and negative pairs among the individual tasks (Table 3). Note however, that the difference (Δ) between positive and negative examples in IE and IR is identical while the accuracy of the word overlap feature differs drastically.

[2] One indicator for the "difficulty" of a test set is the average lexical overlap of text and hypothesis. The difference of the proportion between the entailed and not entailed pairs – the discriminative power of the overlap feature – differs among different sets: e.g. 0.05 on T2 and 0.13 on D3.

[3] In terms of machine learning, extending a training set by factor 2 (from 800 to 1.600 items) does not make a qualitative difference. The improvement observed by (Hickl et al., 2006) was achieved by going to 10.000 items.

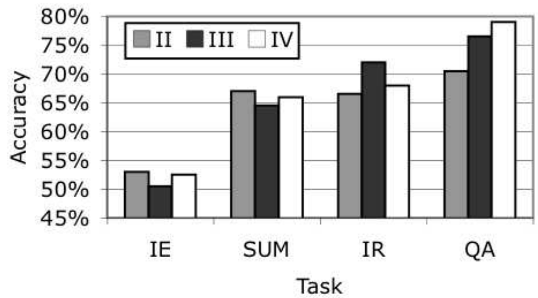

Figure 2: Per task accuracy (D2T2-D3).

Task	Entailed	Not Entailed	Δ
IE	0.41	0.35	0.06
SUM	0.39	0.22	0.17
IR	0.29	0.23	0.06
QA	0.44	0.20	0.24

Table 3: Average word overlap per task for D3

The per task analysis also confirms the observation that the combination of (deep and shallow) features behaves heterogeneously in terms of accuracy.

A somewhat unexpected result is that the more "informed" feature II performs better in SUM as the shallow feature IV while it is the other way round in QA (see Figure 2).

4 Discussion

It is a bit surprising that a shallow feature like word overlap performs comparable to, or even better than, more informed features obtained from a relatively deep linguistic analysis, and that the combination of both types of features does not always increase the overall accuracy.

One possible explanation is the limited size of the training data, which seems to be too small for the machine learner to exploit the full potential of the deep features. One way to compensate for the limited training size would be to make the implicit linguistic information encoded in the features more explicit, for instance by making the graph matching linguistically more informed. Options are to compute a proper embedding of the hypothesis graph into the text graph, or interlinking of the various layers of analysis (syntax, semantics, ontology) in some other more controlled way.

However, to come up with a more explicit model of textual entailment, a deeper understanding of the principles involved in establishing textual entailment relations is necessary. An idea which is derived from the traditional notion of logical entailment is that the information encoded in the hypothesis must (somehow) be subsumed by the text for the entailemt to hold. Although most approaches to textual entailment seem to rely on this assumption in one way or another, it is easy to find pairs in the RTE corpora where the relation between text and hypothesis cannot be modelled so straightforwardly. In (1), for instance, textual entailment holds although *was born in* is more specific than *be from*.

(1a) As a real native Detroiter, I want to remind everyone that Madonna **is from** Bay City, Mich., [. . .].

(1b) Madonna **was born** in Bay City, Mich.

Interestingly, textual entailment sometimes does not hold even if the information expressed by the hypothesis is subsumed by the text:

(2a) [. . .] Nizar Hamdoun, announced today, Sunday, that thousands of people **were killed** or injured during the four days of air bombardment against Iraq.

(2b) Nizar HAMDOON, Iraqi ambassador to the United Nations, announced that thousands of people **could be killed** or wounded due to the aerial bombardment of Iraq.

Although the hypothesis is logically entailed by the text (if we ignore the report context) – 'kill' implies 'possibly kill' – pragmatic principles seem to block entailment here.

The observation that standard logical entailment and textual entailment deviate in certain respects is not surprising and has also been addressed in a discussion initiated by (Zaenen et al., 2005). Still, there is no consensus regarding the precise mechanisms involved in the latter such as "general principles of plausibility" or pragmatic principles.

We conducted a short annotation experiment during a reading circle at our department on a randomly chosen subset of 10 pairs from the RTE 1 (including (1) and (2) from above). A central result was that it is relatively easy to decide *whether* textual entailment holds while it often remained controversial *why* this is the case. In particular, it seems difficult to tell whether an inference is strict or just plausible, and whether it relies on lexical knowledge only or whether "world knowledge" is involved. Currently, a larger subset of the RTE datasets is annotated as part of a Master's thesis project, and we hope to learn more about the principles that underly the notion of textual entailment from the analysis of this data.

5 Conclusion and Outlook

In this paper, we have compared two approximations to textual entailment – a shallow one based on word overlap, and a more informed one based on a relatively deep linguistic analysis. The evaluation on various data sets shows that both perform (by and large) comparable; sometimes the shallow component even outperforms the deeper one. A modest improvement in accuracy can be achieved by combining both components, but this effect cannot be observed invariably on all data sets.

One reason why the deep system does not perform better seems to be the limited size of the training data available for the machine learning component. As we cannot expect the necessary amount of training data to be available in the near future, we currently investigate the data more closely in order to arrive at a more controlled model of textual entailment. In another current effort, we work on an interface to upper-level ontologies (Reiter, 2007) in order to access more "world-knowledge" which is a desideratum in natural language processing in general, as in many approaches to textual entailment.

Acknowledgements

This work has partly been funded by the German Research Foundation DFG (grant PI 154/9-2). We also acknowledge Katrin Erk's support of the Shalmaneser system.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL*, Montreal, Canada.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor, editors. 2006. *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

Johan Bos and Katja Markert. 2006. When logical inference helps determining textual entailment (and when it doesn't). In *Proceedings of PASCAL RTE2 Workshop*.

Aljoscha Burchardt and Anette Frank. 2006. Approximating Textual Entailment with LFG and FrameNet Frames. In *Proceedings of PASCAL RTE2 Workshop*.

Aljoscha Burchardt, Katrin Erk, and Anette Frank. 2005. A WordNet Detour to FrameNet. In B. Fisseni, H.-C. Schmitz, B. Schröder, and P. Wagner, editors, *Sprachtechnologie, mobile Kommunikation und linguistische Resourcen*, volume 8 of *Computer Studies in Language and Speech*. Peter Lang, Frankfurt/Main.

Katrin Erk and Sebastian Pado. 2006. Shalmaneser - a flexible toolbox for semantic role assignment. In *Proceedings of LREC 2006*, Genoa, Italy.

Christiane Fellbaum. 1997. English verbs as semantic net. In *WordNet: an electronic lexical database*. MIT.

Andrew Hickl, Jeremy Bensley, John Williams, Kirk Roberts, Bryan Rink, and Ying Shi. 2006. Recognizing Textual Entailment with LCC's Groundhog System. In *Proceedings of PASCAL RTE2 Workshop*.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS)*, pages 2–9. ACM Press.

Nils Reiter. 2007. Towards linking FrameNet and SUMO. Diploma Thesis (in preparation).

Stefan Riezler, Tracy H. King, Ronald M. Kaplan, Richard Crouch, John T. III Maxwell, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In *Proceedings of ACL'02*, Philadelphia, PA.

Ulrich Schäfer, 2005. *Heart of Gold, User and Developer Documentation*. DFKI Language Technology Lab, Saarbrücken, Germany.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2 edition.

Annie Zaenen, Lauri Karttunen, and Richard Crouch. 2005. Local textual inference: Can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, Ann Arbor, Michigan, June.