Approaching Textual Entailment with LFG and FrameNet Frames

Aljoscha Burchardt
Dept. of Computational Linguistics
Saarland University
Saarbrücken, Germany
albu@coli.uni-sb.de

Anette Frank
Dept. of Computational Linguistics
Saarland University &
Language Technology Lab, DFKI GmbH
Saarbrücken, Germany
frank@coli.uni-sb.de

Abstract

We present a baseline system for modeling textual entailment that combines deep syntactic analysis with structured lexical meaning descriptions in the FrameNet paradigm. Textual entailment is approximated by degrees of structural and semantic overlap of text and hypothesis, which we measure in a *match graph*. The encoded measures of similarity are processed in a machine learning setting.[1]

1 Introduction

In this paper, we present a baseline system to approach the textual entailment task as presented in the PASCAL RTE Challenge. This task faces us with the problem of modeling deep text understanding and inference for complex examples in unrestricted domains. Similar to previous work (Dagan et al., 2005) we explore semantically informed *approximations* of textual entailment. As shown by (Bos and Markert, 2005), fine-grained semantic analysis and reasoning models can yield high precision, but are severely restricted in recall. The architecture we present is open for extension to deeper methods.

We assess the utility of approximating entailment in terms of structural and semantic overlap of text and hypothesis, combining wide-coverage LFG parsing with *frame semantics*, to project a coarse-grained lexical semantic representation with semantic roles. We compute various measures of overlap and train a machine learning model for entailment.

In Section 2, we describe the linguistic resources and our system architecture. In Section 3, we present our approach for modeling similarity of text and hypothesis in a *match graph*. In Section 4, we report on our machine learning experiments, the results in the RTE task, and provide some error analysis, including discussion of typical examples that show the strength and weaknesses of our approach. We conclude with a discussion of perspectives.

2 Base Components and Architecture

2.1 Basic Analysis Components

Our primary linguistic analysis components are the probabilistic LFG grammar for English developed at Parc (Riezler et al., 2002), and a combination of systems developed in the SALSA project: two probabilistic systems for frame and role annotation, *Fred* and *Rosy* (Erk and Pado, 2006) and a rule-based system for frame assignment, called *Detour (to FrameNet)* (Burchardt et al., 2005), which uses WordNet to address coverage problems in the current FrameNet data. In addition we use the Word Sense Disambiguation system (Banerjee and Pedersen., 2003) and mappings from WordNet to SUMO (Niles and Pease, 2003) to assign WordNet synsets and SUMO ontological classes to main predicates.

2.2 Frame Semantics

Frame Semantics (Baker et al., 1998) models the lexical meaning of predicates and their argument

[1]This work has been carried out in the project SALSA, funded by the German Science Foundation DFG, Title PI 154/9-2. We thank Katrin Erk and Sebastian Pado for providing and supporting the Fred and Rosy systems and Alexander Koller for his contributions and for implementing the FEFViewer.

Role	Example
SELLER	*BMW bought Rover from **British Aerospace**.*
BUYER	*Rover was bought by **BMW**, which financed [. . .] the new Range Rover.*
GOODS	*BMW, which acquired **Rover** in 1994, is now dismantling the company.*
MONEY	*BMW's purchase of Rover for **$1.2 billion** was a good move.*

Figure 1: Frame COMMERCE_GOODS-TRANSFER.

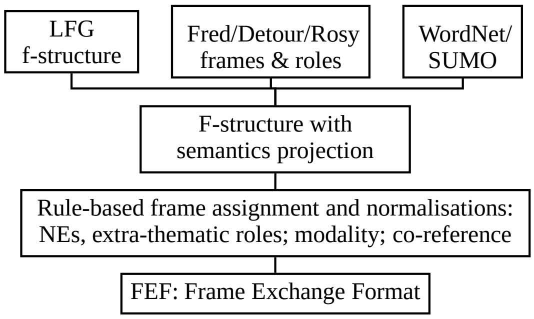

Figure 2: Architecture of linguistic analysis

structure in terms of *frames* and *roles*. A *frame* describes a conceptual structure or prototypical situation together with a set of *semantic roles* that identify participants involved in the situation. FrameNet currently contains more than 600 frames with almost 9000 lexicalizations (word-frame pairs). Figure 1 displays examples involving the frame COMMERCE_GOODS-TRANSFER.

Frame-semantic analysis is especially interesting for the task of recognizing textual entailment if we aim at robust and high-quality measures for semantic overlap. Frames provide normalisations for diverse surface realizations (lexicalisation, verb vs. nominalisation, etc.), including variations in argument structure realisation (cf. Fig. 1). Thus, we can determine *semantic similarity based on lexical semantic meaning*, combined with measuring *similarity of argument structure* at a high level of abstraction. Moreover, the coarse-grained frame structures make it possible to assess the core meaning of a sentence ("what is it about?") in a shallow analysis, separated from the pitfalls of deep, structural analysis of scope, modality, etc.,which must be treated by other components, or can be selectively introduced, as will be illustrated for the case of modality.

2.3 Enriched Frame Semantic Representations

As displayed in Figure 2, LFG-based syntactic analysis (i.e., f-structure) is integrated with frames and roles assigned by Fred, Detour and Rosy, as well as WordNet synsets and SUMO concepts, to yield an f-structure with frame-semantic projection (Frank and Erk, 2004), including conceptual class assignments. The integration and semantics projection is defined using the XLE rewrite system of (Crouch, 2005).

Additional rules introduce frames and concept classes based on *named entities* recognized in LFG parsing (companies, political offices etc.), as well as *extrathematic semantic roles* (TIME, LOCATION, REASON, etc.) for corresponding adjunct types in f-structure. We also collect possible *antecedent referents* for pronominals, as a heuristic device to establish co-referential links. Finally, we identify various types of *modal contexts*, such as negation, modals, conditionals or future tense that allow to detect text-hypothesis pairs that preclude entailment.

The result structures are converted to a *Frame Exchange Format (FEF)*, a flat predicate representation comprising syntactic and semantic analysis. Table 1 displays the FEF for (1). The parts printed in bold show information from different levels for the predicate *manufacturer*: f-structure node f(5), semantics projection to node s(61) which is labled with the frame MANUFACTURING (with roles PRODUCT and MANUFACTURER) plus a projection to ontological information (s(71)), WordNet synset and SUMO super-class in this case. A FEFViewer (Figure 3) displays the major syntactic and semantic graph structures.

(1) *Mercedes-Benz is a German car manufacturer*.

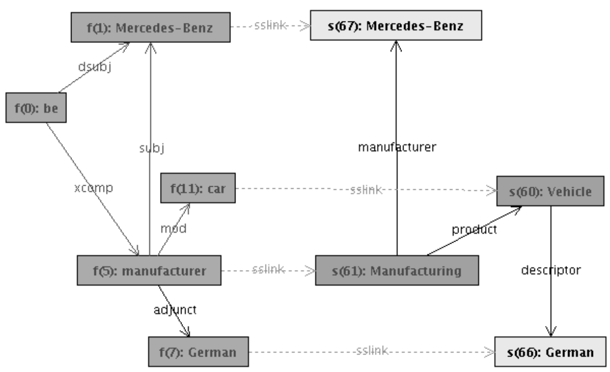

Figure 3: FEFView for example (1).

normalized f-structure with syn-sem projections	frames, roles and ontological info (WordNet/SUMO)
xcomp(f(0),f(5)).	frame(s(60),'Vehicle').
tense(f(0),pres).	vehicle(s(60),s(60)).
stmt_type(f(0),declarative).	descriptor(s(60),s(66)).
pred(f(0),be).	rel(s(66),'German').
mood(f(0),indicative).	
dsubj(f(0),f(1)).	**frame(s(61),'Manufacturing').**
pred(f(1),'Mercedes-	product(s(61),s(60)).
Benz').	manufacturer(s(61),s(67)).
num(f(1),sg).	rel(s(67),'Mercedes-
subj(f(5),f(1)).	Benz').
pred(f(5),manufacturer).	
num(f(5),sg).	ont(s(60),s(72)).
mod(f(5),f(11)).	ont(s(66),s(73)).
det_type(f(5),indef).	**ont(s(61),s(71)).**
adjunct(f(5),f(7)).	
pred(f(7),'German').	**wn_syn(s(71),'manufacturer#1').**
atype(f(7),attributive).	**sumo_sub(s(71),'Corporation').**
adjunct_type(f(7),nominal).	milo_sub(s(71),'Corporation').
adegree(f(7),positive).	
pred(f(11),car).	wn_syn(s(72),'car#n#1').
num(f(11),sg).	sumo_sub(s(72),'Transp~Device').
	milo_sub(s(72),'Transp~Device').
sslink(f(1),s(67)).	
sslink(f(5),s(61)).	wn_syn(s(73),'german#a#1').
sslink(f(7),s(66)).	sumo_inst(s(73),'Nation').
sslink(f(11),s(60)).	milo_syn(s(73),'Germany').

Table 1: FEF for example (1).

2.4 Overall RTE Architecture

Our RTE system architecture comprises the folowing steps: We compute LFG f-structures with extended frame semantics projections for text and hypothesis pairs. We identify their structural and semantic similarities and represent them in a *match graph*. From text, hypothesis, and match graph we extract features that characterize their syntactic and semantic properties, as well as various relational properties that can be considered relevant for establishing or rejecting entailment. These features are fed into a Machine Learning system for training on the development set and testing on the test set.

3 Computing Semantic Overlap

We approximate textual entailment by statistical prediction on the basis of measurements for structural and semantic overlap between text and hypothesis.

3.1 Matching Text and Hypothesis

In a graph matching process we compute the overlap of the f-structures with semantics projection (i.e. graphs) for text and hypothesis which we record in a *match graph*. The latter consists of matched *predicates* and *features* from both input graphs. We distinguish various (sub)types of matches, in order to selectively extract features for the learning phase.

Node (predicate) matching. Node matching rules match nodes for *identical* syntactic predicates and frames. We also allow matches for predicates that are semantically related on the basis of *WordNet*. To prevent overgeneration, WordNet-based matching is restricted to predicates that are related by an edge in the match graph. Further, the respective synsets have to be closely related in terms of WordNet path distance (<3). Using (heuristically defined) antecedent sets for pronouns, we allow special types of predicate matches for pronouns and non-pronominal predicates in text and hypothesis.

In addition, we allow matches between frame nodes that are known to be related by *FrameNet frame relations*, such as *inheritance*, or those that are considered related by the Detour system, measuring *frame distance* on the basis of WordNet.

Feature (edge) matching. Feature matches are restricted to features that connect matching nodes, or those that take identical atomic values. The linguistic nature of these edges ranges from morphosyntactic features in LFG f-structure, such as NUM, PERS, over grammatical functions ((deep) subject, (deep) object, adjunct, oblique, complement, etc.), to frame semantic roles in the semantic projection.

Modality contexts Besides finding matches for *similar* nodes and edges, some rules are intended to detect *semantic difference* in terms of incompatible modality types. We normalise the different modal contexts to five basic types: conditional, subjunctive, diamond, box and negation. An example of incompatible modalities is the pair: *A pet must have rabies protection confirmed by a blood test – A case of rabies was confirmed*.

3.2 Feature Extraction

The features we extract from the text, hypothesis and match graphs to induce a machine learning model for textual entailment can be classified according to their (i) *nature* in terms of level of representation (lexical, syntactic, semantic), (ii) *degree of connectedness* in matching, (iii) *source* (text, hypothesis

1.	No. of predicate matches relative to hypothesis.
2.	No. of frame (Fred, Detour) matches relative to hypothesis.
3.	No. of roles (Rosy) matches relative to hypothesis.
4.	Match graph size relative to hypothesis, including syntactic, semantic, and ontological information.

Table 2: Feature Set for Submitted Test Runs

	All tasks	IE	IR	QA	SUM
run1	0.59	0.50	0.60	0.55	0.73
run2	0.58	0.49	0.59	0.57	0.67

Table 3: RTE 2006 results: Accuracy.

or match graph), and (iv) *proportional relation* (hypothesis/text, match-graph/hypothesis ratio, etc.).

Lexical features count the number of lexical items, syntactic features record the number of LFG *predicate* matches, including pronominal and coreferential matches in the match graph, and *syntactic features*. Semantic features distinguish between those frames and roles that were assigned by the Fred, Detour and Rosy systems, and those that were successfully interfaced with LFG analyses.[2] We further distinguish semantic node matches of different types as discussed above (e.g. identical or semantically related frames, modal properties). Finally, we compute the number and size of connected clusters in the match graph, as well as the relative size in relation to the size of the hypothesis graph.

4 Experiments and Results

4.1 Training and Classification

Feature selection. We experimented with various learners and the attribute selection module of Weka (Witten and Frank, 2005). A general observation was that many learners (evaluators) select features that seem intuitively important. However, also unintuitive features, such as the frequency of predicates in the hypothesis graph, showed up as high-valued features, which could be due to idiosyncrasies in the development set. We chose to submit a run that is based on a small and intuitively plausible feature set which led to constant results on a number of classifiers. The feature set is listed in Table 2.

Results. We submitted two runs for different classifiers from Weka, using the feature set from Table 2. For run1, we used a simple conjunctive rule classifier. It generated a single rule measuring predicate and frame matches relative to the hypothesis:

$$(\text{preds\_m\_relto\_h} \leq 0.485294) \text{ and } (\text{frames\_m\_relto\_h} \leq 0.954546) \Rightarrow \text{rte\_entails} = 0$$

For run 2, we used the *LogitBoost*[3] classifier from Weka's meta classifers which used all features, except for role assignments, in its iteration steps. The official RTE results are listed in Table 3.

4.2 Discussion of Results and Error Analysis

The conjunctive rule used in run1 imposes a medium and high threshold, respectively, on predicate and frame matches, as criteria for rejection. So, the system accepts high degrees of semantic similarity based on frames, joint with medium degree overlap at the syntactic predicate level to model entailment.

This is in accordance with the view that frame semantics models "aboutness", on the basis of coarse-grained conceptual meaning, as opposed to veridicality as it is modeled by truth-conditional semantics. This is further confirmed by the results for the different RTE tasks (Table 3): we obtain higher accuracy for SUM (and IR), as opposed to QA and IE, which (in the RTE setting) need deeper modeling in terms of veridicality. Run 2, which uses the more "informative" feature set of Table 2 performs only slighly worse than run 1, and better on QA.

True positives. Table 4 lists examples of true positives. Entailment is triggered by high semantic overlap between hypothesis and match graph in terms of matching predicates, frames, and f-structure. Ex. 602 is an example where frames establishes a semantic match for predicates without a syntactic match: the verb *purchase* and the nominal *purchase* are both assigned the frame COMMERCE_BUY.

On the other hand, missing or non-matching frame assignments can be compensated via WordNet relatedness: in ex. 103, *die* is matched with *pass away* although the latter has not been assigned a frame. Active-passive diathesis such as *soldier*

[2] A number of frames and roles could not be ported from Fred and Detour onto the f-structure due to mismatches in lemmatisation/tokenisation and fragmentary or failed parses.

[3] LogitBoost performs *additive logistic regression* using the classifier *DecisionStump*.

True positives:

103 T: *Everest summiter David Hiddleston has passed away in an avalanche of Mt. Tasman.*
H: *A person died in an avalanche.*

129 T: *In one of the latest attacks, a US soldier on patrol was killed by a single shot from a sniper in northern Baghdad, the military said yesterday.*
H: *A sniper killed a U.S. soldier on patrol in Baghdad with a single shot.*

602 T: *The system of government purchases of food under the U.N. Oil-for-Food Program was alleged to have many abuses.*
H: *A government purchases food.*

626 T: *An earthquake has hit the east coast of Hokkaido, Japan, with a magnitude of 7.0 Mw.*
H: *An earthquake occurred on the east coast of Hokkaido, Japan.*

True negatives:

233 T: *The goal of preserving indigenous culture* **can** *hardly be achieved by a handful of researchers and curators at museums of ethnology and folk culture.*
H: *Indigenous folk art is preserved.*

322 T: *Even today, within the deepest recesses of our mind, lies a primordial fear that* **will** *not allow us to enter the sea without thinking about the possibility of being attacked by a shark.*
H: *A shark attacked a human being.*

Table 4: Examples from RTE 2006.

was killed and *killed a soldier* in ex. 129 is resolved on the f-structure level where we normalize to deep subject and object. As seen in ex. 626 and 129, good results are not only obtained for short hypotheses.

True negatives. 27% of justified rejections involve mismatches of modality, while only 11.9% of all sentences contain modal contexts. Our matching algorithm for construction of the match graph includes a heuristics that rejects predicate (and feature) matches if the predicates (features) are embedded in inconsistent modal contexts. Thus, mismatching modalites are reflected in two ways: by (distinct) modality features in text and hypothesis, and in terms of reduced size of the match graph. Ex. 233 and 322 are true negatives where predicate matches of the underlined predicates are blocked.

Error analysis for base components. LFG parsing yielded 99% coverage for the test set. 24% of the sentence *pairs* involved a fragmentary parse. For these, we rely on non-LFG-integrated frame and role assignments by Fred, Rosy and Detour. To assess the impact of losses in syntactic analysis, enriched semantic representations and the resulting overlap measures, we restricted the test set to pairs without fragmentary parses, which yielded an improvement of 1-3% for various learners and feature sets.

Overall, the system assigned 14326 frames and 13325 roles, including 3199 frames and 1736 roles added by default rules. In average, 8,9 frames per sentence and 1.1 role per frame. We identified losses in the interface that projects frames and roles to the LFG (10% for frames, 38,9% for roles) that are due to failed or partial parses, but also to remaining differences in tokenisation and lemmatisation. Losses in porting frame and role assignments to LFG are compensated by the fall-back to non-assigned frames and roles, though they do have an impact on the computation of the overlap features, such as connectedness and size of the match graph.

Sparse features. From a machine learning view, the size of the development corpus is very small. Phenomena (features) that do not occur in the majority of sentence pairs are neglected by the machine learning systems. Currently, we have high-frequency features that measure *similarity* (e.g. predicate and frame overlap), but only few and low-frequency features that identify *dissmimilarity*, such as mismatching modalities. Therefore, the learners have a tendency to reject too little: 29,5% false positives as opposed to 12,75% false negatives.

False positives and negatives. False positives often involve non-matching main predicates that are in fact semantically dissimilar within larger match graphs. In line with the above observation of sparse features for dissimilarity, we see potential for improvement by including additional measures for semantic distance between non-matching nodes in otherwise connected match graphs.

A related problem we observed for nodes in the match graph that are closely connected e.g in the hypothesis, but come from far distant parts of the text graph, as in ex.198: ***4.4 million people** were **executed** in Singapore – Some 420 people have been **hanged** in Singapore[...]. That gives the country of **4.4 million people** the highest execution rate.*. For such configurations, we could establish a new type of *weighted* edge match that reflects the relative distance of the node pairs in the text and hypothesis graphs, measured in terms of f-structure or frame structure path distance. This, we hope, could help the learner to establish further criteria for rejection.

Inferences on partial structures. Our architecture is open for extension to deeper methods. We have started to integrate *inferences* on partial structures in order to bridge partial non-matching text and hypothesis graphs: e.g., $joins(x_1, y_1)$ in the text graph supports the hypothesis $member\_of(x_2, y_2)$, for matching node pairs (x_1/x_2, y_1/y_2). In the graph matching process, inferences of this type introduce special types of matches, which can be exploited by the learner directly, or indirectly, through the ensuing extension of the match graph. However, due to the small, manually crafted rule set, this feature was not yet effective. The next step is thus to identify and integrate suitable, large-scale resources for inferences, both lexical and based on world-knowledge.

5 Conclusions and Perspectives

We presented a baseline system for textual entailment that is based on "informed" features for structural and semantic overlap between text and hypothesis. The system's performance is on a par with the best systems in last year's RTE Challenge. We consider this to demonstrate the usefulness of a frame-based approach to textual entailment – combined with deep syntactic analysis and further components that complement aspects of semantic modelling not covered in frame semantics.

We identified various possibilities for further improvement. The current bias towards positive entailment judgments can be compensated by introducing more *negative* features that measure the *distance* – semantic or constructional – between material involved in partial match graphs. More generally, starting from the determination of *structural and semantic overlap*, or *similarity*, we can now improve the modelling of *dissimilarity*. The detection of *incompatible modalities* has proved rather effective, but can be further extended to *lexically induced* modalities (e.g. *possibility of, alleged, promise*).

The usage of an integrated syntactic-semantic-ontological representation supports the integration of selected deeper and fine-grained methods for semantic analysis, in terms of measures for similarity, dissimilarity, or inferences on partial structures.

References

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

Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, Acapulco, Mexico.

Johan Bos and Katja Markert. 2005. Combining shallow and deep NLP methods for recognizing textual entailment. In *Proceedings of the First Challenge Workshop, Recognizing Textual Entailment*. PASCAL.

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*, pages 408–421. Peter Lang, Frankfurt am Main.

Richard Crouch. 2005. Packed Rewriting for Mapping Semantics to KR. In *Proceedings of the Sixth International Workshop on Computational Semantics, IWCS-06*, Tilburg.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Proceedings of the First Challenge Workshop, Recognizing Textual Entailment*. PASCAL.

Katrin Erk and Sebastian Pado. 2006. Shalmaneser – a toolchain for shallow semantic parsing. In *Proceedings of LREC-2006 (to appear)*, Genoa, Italy.

Anette Frank and Katrin Erk. 2004. Towards an LFG Syntax-Semantics Interface for Frame Semantics Annotation. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, LNCS, pages 1–12. Springer.

Ian Niles and Adam Pease. 2003. Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In H.R. Arabnia, editor, *IKE*. CSREA Press.

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.

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