# *FATE*: a FrameNet-Annotated corpus for Textual Entailment

**Aljoscha Burchardt, Marco Pennacchiotti**

Dept. of Computational Linguistics
Saarland University
Saarbrücken, Germany
{albu,pennacchiotti}@coli-uni.sb.de

**Abstract**

Several studies indicate that the level of predicate-argument structure is relevant for modeling prevalent phenomena in current textual entailment corpora. Although large resources like FrameNet have recently become available, attempts to integrate this type of information into a system for textual entailment did not confirm the expected gain in performance. The reasons for this are not fully obvious; candidates include FrameNet's restricted coverage, limitations of semantic parsers, or insufficient modeling of FrameNet information. To enable further insight on this issue, in this paper we present **FATE** (**F**rameNet-**A**nnotated **T**extual **E**ntailment), a manually crafted, fully reliable frame-annotated RTE corpus. The annotation has been carried out over the 800 pairs of the RTE-2 test set. This dataset offers a safe basis for RTE systems to experiment, and enables researchers to develop clearer ideas on how to effectively integrate frame knowledge in semantic inferenence tasks like recognizing textual entailment. We describe and present statistics over the adopted annotation, which introduces a new schema based on full-text annotation of so called *relevant* frame evoking elements.

## 1. Introduction

It is a commonplace that semantic knowledge plays an important role in Natural Language Processing, especially in view of the challenge of providing user-friendly information access to huge textual corpora like the World Wide Web. Yet, current approaches to information access mostly neglect semantic knowledge.

The Recognizing Textual Entailment (RTE) task (Dagan et al., 2006; Bar-Haim et al., 2006; Sekine et al., 2007) offers a suitable semantic framework to study the role of semantic knowledge in information access applications. Indeed, RTE subsumes most inference based tasks, such as Question Answering, Information Retrieval and Information Extraction. The RTE scheme is straightforward – two sentences called the *text* (T) and the *hypothesis* (H) are said to stand in a textual entailment relation if a typical language user would say that H follows from T, as in the following example.

(1) T: Yahoo has recently acquired Overture.

(2) H: Yahoo owns Overture.

So far, various methods have been used for RTE, but it is not yet clear (i) to what extend and how different semantic resources can effectively contribute and (ii) how actual systems can make optimal use of existing resources (e.g., find the best feature model in a machine learning system). In fact, results of the past three years' RTE challenges (e.g., Bar-Haim et al. (2006)) show that shallow distributional methods using little semantics (e.g., only WordNet) still tend to outperform "deeper" semantic methods (e.g., Bos and Markert (2005), Burchardt et al. (2007)).

In this paper, we will focus on the contribution of lexical semantic knowledge at the level of predicate-argument structure. Several studies (e.g., Bar-Haim et al. (2005), Litkowski (2006)) indicate that this level of granularity is relevant for modeling many phenomena which occur in the current textual entailment corpora, such as lexical alternations, variations and paraphrases. Resources at the predicate-argument level could then play a central role for supporting RTE systems. To date, two major resources are available: PropBank (Kingsbury et al., 2002) and FrameNet (Baker et al., 1998). PropBank models variation only within predicates. FrameNet, on the other hand, abstracts over individual predicates and groups words evoking the same situation type into frames, thus modeling relations among different predicates and parts of speech. FrameNet should then offers a better and wider support to RTE.

Still, a positive impact of FrameNet on the task of RTE has not been confirmed. In fact, the only existing RTE system based on FrameNet (Burchardt and Frank, 2006) performed only at the middle ranges. The reasons of this limited impact are still not clear, the most plausible being: coverage issues of FrameNet; limited reliability of frame semantic parsers; not optimal use of the frame semantic information in the reasoning component. In order to fully leverage predicate-argument knowledge in tasks such as RTE, it is necessary to understand which of these is the main limiting factor.

In this paper we present **FATE** (**F**rameNet-**A**nnotated **T**extual **E**ntailment), a manually crafted, fully reliable frame-annotated RTE corpus. FATE consists of the 800 $(T, H)$ entailment pairs from the RTE-2 Challenge test set, annotated with frame and semantic role labels. The main goal of our annotation effort is to give a practical help to disentangle the above problem. Indeed, our dataset contributes: (i) evidence as to whether FrameNet coverage over the RTE corpora is sufficient to allow inference at the predicate-argument level; (ii) a gold standard for testing the performance of existing shallow semantic parsers on realistic data; (iii) a basis that enables researchers to develop clearer ideas on how to effectively integrate frame knowledge in semantic inference tasks like RTE; (iv) a noise-free frame-annotated corpus for RTE systems to experiment on.

The paper is structured as follows. In section 2., we provide some background on frame semantics and the state-of-the-art in frame-based processing. We also illustrate how frame

semantics can contribute to the task of textual entailment. Section 3. showcases the annotation scheme of FATE, our manual frame semantic annotation of an RTE dataset. In section 4., we discuss the annotation process and provide statistics. Section 5. draws final conclusions and outlines future works.

## 2. FrameNet for RTE inference

The **FrameNet** project provides a collection of linguistically motivated conceptual structures called *frames* that describe prototypical situations. Each frame comes with its own set of semantic roles, called *frame elements* (FEs). These are the participants and propositions in the abstract situation described. From a linguistic perspective, a frame is a semantic class containing predicates that can *evoke* the described situation. These target words or expressions are called *frame evoking elements* (FEE). Table 1 shows the frame STATEMENT, which describes a specific type of a communication situation and is evoked by verbs such as *acknowledge* or *admit*, and by nouns such as *affirmation*.

| Frame: STATEMENT | | |
|---|---|---|
| This frame contains verbs and nouns that communicate the act of a SPEAKER to address a MESSAGE to some ADDRESSEE using language. A number of the words can be used performatively, such as *declare* and *insist*. | | |
| FEs | SPEAKER | **Evelyn** said she wanted to leave. |
| | MESSAGE | Evelyn announced **that she wanted to go**. |
| | ADDRESSEE | Evelyn spoke **to me** about her past. |
| | TOPIC | Evelyn's statement **about her past** |
| | MEDIUM | Evelyn preached to me **over the phone**. |
| FEEs | acknowledge.v, acknowledgment.n, add.v, address.v, admission.n, admit.v, affirm.v, affirmation.n, allegation.n, allege.v, announce.v, announcement.n, assert.v, assertion.n, attest.v, aver.v, avow.v, avowal.n, ... | |

Table 1: Example frame from the FrameNet database.

In the case of STATEMENT, the FEs are the SPEAKER and ADDRESSEE of the statement, the MESSAGE conveyed and its TOPIC. Roles are local to individual frames, thus avoiding the commitment to a small set of universal roles, whose specification has turned out to be infeasible in the past. The current on-line version of the frame database contains about 800 frames and 10.000 lexical entries with annotated example sentences.[1]

**Frame-based processing.** A a freely available state-of-the-art semantic parser is Shalmaneser (Erk and Pado, 2006). It is based on machine learning techniques and pre-trained on the FrameNet corpus. Shalmaneser offers a complete "tool-box" architecture for frame and role assignment, with pre-processing modules to elaborate input from the Collins and the Minipar syntactic parsers. Shalmaneser offers high performance, with an accuracy of 0.93 on frame assignment, and F-scores of 0.85 and 0.78 respectively on role recognition and labeling (Pado, 2007) if evaluated on the FrameNet sample corpus. Shalmaneser can be boosted with the "Detour to FrameNet" (Burchardt et al., 2005), a rule-based frame assignment system, which addresses lacks in FrameNet's coverage by using WordNet to infer correct frame assignment for unknown FEEs. Shalmaneser and Detour have in fact been used in combination in the frame-based RTE system of Burchardt and Frank (2006).

[1] http://framenet.icsi.berkeley.edu

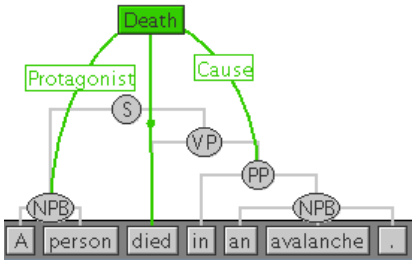


Figure 1: Frame semantic analysis of (4).

**RTE-2 dataset.** The FATE annotation is built over the RTE-2 challenge test set corpus. This corpus consists of 800 $(T, H)$ pairs, similar to the one reported in the Introduction. Pairs are created using both automatic and supervised techniques inspired by common NLP tasks: Question Answering, Information Extraction and Multi Document Summarization. The dataset was then annotated by two human judged, which had to classify a pair to be either a positive or a negative example of textual entailment. The resulting inter-annotator agreement was 0.78, corresponding to *substantial agreement*. All pairs in disagreement were discarded, and a further check was finally done by a third judge. A full description of the dataset and on its building procedure is presented in (Bar-Haim et al., 2006). Throughout the paper, we will show several entailment pair examples.

**Frames for modeling textual entailment.** The annotation of predicate-argument structure in general, and of frames in particular, is interesting for its intermediate position between syntax and "deep", compositional semantics. Frame semantics disregards problems of deep semantic analysis such as modality, negation, or scope ambiguity and instead structures meaning information on the level of aboutness ("who did what to whom"). This level of granularity is attractive for modeling many phenomena occurring in the currently available textual entailment corpora. As illustration, consider the sentence pair below from the RTE-2 corpus (Bar-Haim et al., 2006).

(3) T: [Everest summiter David Hiddleston]$_{\text{PROTAGONIST}}$ has passed away [in an avalanche of Mt. Tasman]$_{\text{CAUSE}}$. (frame: DEATH)

(4) H: [A person]$_{\text{PROTAGONIST}}$ died [in an avalanche]$_{\text{CAUSE}}$. (frame: DEATH)

Figure 1 shows a graphical representation of the frame annotation of the hypothesis on top of a syntactic parse provided by the Collins parser (Collins, 1999). The frame DEATH is evoked by the verb *died*, the PROTAGONIST role points to *a person*, the CAUSE role to *in an avalanche*.

The frame annotation for the text (3) is quite similar. The phrasal verb *pass away* also evokes the frame DEATH, the PROTAGONIST role points to *Everest summiter David Hiddleston*, the CAUSE role to *in an avalanche of Mt. Tasman*. Evidently, the frame analysis provides a semantic normalization – it shows that both sentences talk about the same situation and participants. This is a strong evidence for an entailment relation. The last bit of information needed to confirm that textual entailment actually holds, namely testing whether *person* and *David Hiddleton* are compatible and likewise *avalanche* and *avalanche of Mt. Tasman*, does not fall into the realm of frame semantics. This can be done in subsequent processing steps using other means and resources, e.g., string comparison, named entity recognition and thesauri.

Likewise, frame semantics generalizes across near meaning-preserving transformations such as argument variation, alternation in voice and word class or in lexicalization (e.g. *"Evelyn spoke about her past"* vs *"Evelyn's statement about her past"*). FrameNet can also account for not so straightforward, inferential relations via the existing frame hierarchy. Consider (5)/(6) from the RTE-3 development corpus.

(5) T: El-Nashar was detained July 14 in Cairo. Britain notified Egyptian authorities that it suspected he may have had links to some of the attackers.

(6) H: El-Nashar was arrested in Egypt.

As can be seen in Figure 2, the main verbs of both sentences evoke different frames, respectively DETAINING and ARREST. Also, the roles are slightly different (HOLDING_LOCATION vs. PLACE). Yet, both frame inherit from a common ancestor, INHIBIT_MOVEMENT. As frame inheritance also includes the roles, it is possible to come up with a uniform analysis of both sentences. Again, the information that *Cairo* and *Egypt* are in fact compatible has to be provided by other sources.

As we mentioned in the introduction, several studies confirm the intuition that the level of granularity offered by FrameNet is relevant for modeling many phenomena which occur in the current textual entailment corpora. For example, (Bar-Haim et al., 2005) show that 31% of the RTE-2 positive dataset involves paraphrase at the predicate level. These numbers are comparable to those obtained in the RTE-2 *ARTE* annotation (Garoufi (2007), see section 3.), which demonstrates that at least 20% of the positive examples in the RTE-2 test set can be treated by inferences at the frame level (such as nominalizations and argument variations).

## 3. Annotation scheme

In the literature, FrameNet based corpus annotation follows two basic schemata: *lexicographic annotation*, where only selected, representative predicates from a reference corpus are annotated, and *full-text annotation*, where a whole text or corpus is completely annotated. For the task at hand of annotating a textual entailment corpus, the latter scheme is appropriate.

Within full-text annotation, different strategies have been pursued so far. In the FrameNet project full-text annotation process (Ruppenhofer et al., 2007), annotators work through a given corpus word-by-word. They select any word that can potentially evoke a frame as FEE and annotate it either with an existing frame, or by creating a new one on the fly. In the Salsa corpus annotation (Burchardt et al., 2006a), the annotation has been done predicate-by-predicate. First, all sentences containing a specific FEE are extracted from a reference corpus; then they are annotated with respect to that FEE.

For our general annotation of $(T, H)$ sentence pairs, we follow a slightly modified full-annotation scheme, as we annotate as FEE only *relevant* words (a notion we will make precise below). For the annotation of single FEE instances, we capitalize our annotation experience in the Salsa by adhering to its main guidelines. For example we follow the maximization principle (i.e. when annotating role fillers we chose the largest possible constituent), and the locality principle (i.e. in case of co-reference, we annotate only the local filler). In the rest of this section, we give an overview of the central aspects of our annotation scheme.

**FEE annotation: the relevance principle.** The most critical issue in full-text annotation is to choose which words should be annotated as FEEs. In our context, we want to annotate only words that evoke frames which are somehow *relevant* to the overall situation(s) described in the text at hand. We call such words *relevant FEE*. Indeed, textual entailment inferences are mainly supported by properties and descriptions of relevant facts. Unlike in the FrameNet project annotation, we ask the annotator to skip words evoking a frame which is not central to the situation at hand. The following example illustrates our principle of FEE *relevance annotation* and how it differs from FrameNet annotation. The FrameNet project annotation would select as FEEs all words displayed in boldface below.

(7) T: **Authorities** in Brazil **say** that **more** than 200 **people** are being **held hostage** in a **prison** in the **country's remote**, Amazonian **jungle state** of Rondonia.

(8) H: **Authorities** in Brazil **hold** 200 **people** as **hostage**.

In our annotation schema, we only annotate the relevant FEEs, which are underlined. Indeed, these are the only words which evoke frames describing the overall situations in $H$ and $T$ – *"hostage"* evokes the KIDNAPPING frame, *"say"* evokes STATEMENT.

As described above, the notion of *relevant FEE* has an intuitive flavor, and it may seem to depend mostly on the reader's personal interpretation of the text. To come up with a more operational notion of relevance, that can be applied systematically in the annotation process, we conducted a pilot annotation. We asked two experienced researchers to independently annotate FEEs over the same small set of 15 example sentences randomly extracted from the RTE-2 corpus. The researchers were guided only by the intuition that a good FEE should evoke a relevant situation. Surprisingly, the result showed a very high level of agreement: among

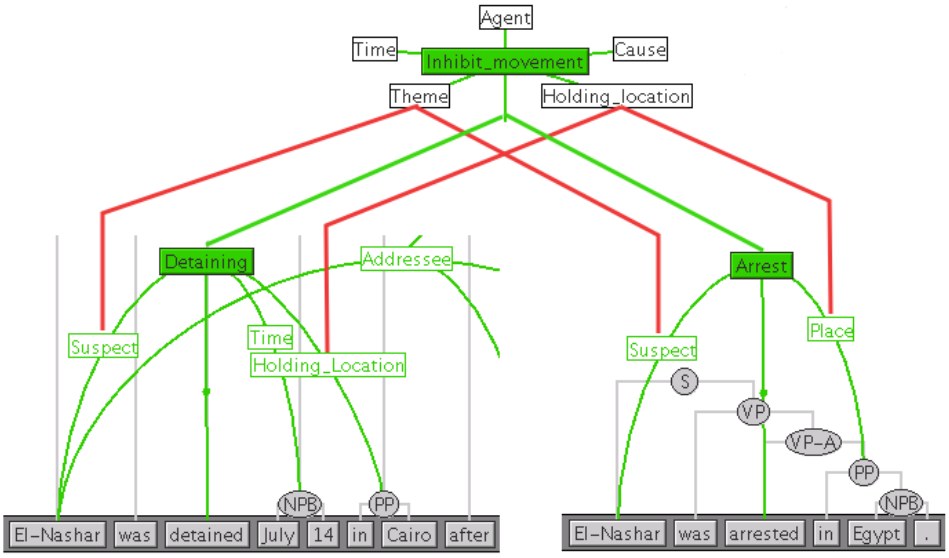


Figure 2: Making Use of Frame Relations.

the 30 relevant FEEs found by the first annotator, and the 32 found by the second, 27 where shared among the two, giving an agreement of 87%. Examination of this preliminary annotation revealed that there are two important properties that help discriminating among relevant and non-relevant FEEs. First, all relevant FEEs have at least one role instantiated in the text. Second, exceptions to this rule are "non-situational" frames like CALENDRIC UNIT and CARDINAL NUMBERS. They are irrelevant although they typically realize roles pointing to the respective numbers.
From this, we adopt the following operational notion of relevance: *a **relevant FEE** is a FEE that evokes a situational frame, and that instantiates in the text at least one role of the evoked frame.*[2]

**Span annotation on positive pairs.** In textual entailment pairs, typically only parts of the texts contribute to the inferential process that allows to derive $H$ from $T$. These cases are most common in positive entailment examples, where the $T$ is composed by one or more long sentences embedding only on a small part the knowledge needed for deriving the entailment. For example, in the following pair, only the sections in bold face are really important:

(9) T: Soon after the EZLN had returned to Chiapas, Congress approved a different version of the COCOPA Law, which did not include the autonomy clauses, claiming they were in contradiction with some constitutional rights (private property and secret voting); this was seen as a betrayal by **the EZLN and other political groups.**

(10) H: **EZLN is a political group.**

[2]Three more guidelines better specify the definition: (1) Cases in which all role fillers are self-references to the FEE must be considered non relevant; (2) In the case that a candidate relevant FEE evokes a situation which is not represented as a frame in FrameNet, the annotator can evoke a special *unknown* frame; (3) A relevant FEE can be either a single word or a multiword expression.

To speed-up the annotation, we decided to annotate only the specific sections within the $(T, H)$ pairs that contain interesting material for the task of textual entailment recognition. The annotators were provided with a markup of the these sections, we call *spans*
We call this *span-annotation*, in contrast with *extensive-annotation* where the full text and hypothesis are annotated. Span-annotation is carried out only on positive examples, as only on these it is possible to clearly point out which sections are interesting, as in the example above. On the contrary, for negative examples the situation is not so clear-cut: in many cases, it is not possible to indicate which portions of texts contribute to a false entailment. Consider for example the following negative pair:

(11) T: Watching Mosaic from the Bay Area, Silicon Graphics CEO Jim Clark, a veteran of the UNIX standards wars, understood how much money could be won if a company could take control of the standards of this new Internet tool.

(12) H: Silicon Graphics created the Internet browser Mosaic.

In the above example, we cannot say that a specific part of $T$ contributes more significantly than another to infer a false entailment.
Spans for the positive $Ts$ are automatically derived by using the ARTE annotation (Garoufi, 2007), which provides alignment annotations for the positive pairs in the RTE-2 test set. The basic observation underlying the ARTE annotation scheme is that if a text entails a hypothesis, then it is usually possible to *embed* the hypothesis into the text. Accordingly, textual entailment is annotated in ARTE by providing mappings (alignments) from so-called *markables* in the hypothesis to markables in the text. Markables are short sequences of words, typically consisting of a single content word plus its dependant function words. The ARTE annotation focuses on properties of the relation between markables in text and hypothesis. It provides a relatively rich set

of features that can be used to annotate the precise properties of the alignments, but what is more important here is that this annotation can be easily used to identify the relevant spans, i.e., spans containing all the lexical material needed to infer the hypothesis from the text, by considering the smallest section of the text which contains all markables used in the alignments.

**Special frames.** We explicitly annotate two pseudo-frames, to cope with the following situations:

- *Unknown frame*. This frame is used when the annotator finds a relevant FEE which evokes a situation not represented in the FrameNet hierarchy. We prefer to adopt an UNKNOWN frame with unknown roles (called *missing FE*) instead of creating explicitly a new frame, because that would imply a specific lexicographic work, which is out of the scope of our annotation process. Statistics on the use of the unknown frame can be leveraged to have an approximation of the FrameNet resource coverage on a RTE dedicated corpus.

- *Anaphora frame*. Anaphoric expressions are widely used in language, and are particularly relevant in textual entailment inference. To comply to the locality principle, we decide to annotate the local referent of an anaphoric role filler, and to link the local referent to the external referent through the ANAPHORA frame. Figure 3 shows an example, where the local filler *"who"* links to the external reference *"former European Commission chief Romano Prodi"* through the anaphora frame. The figure also shows an example of locality principle: the filler of the ANTECEDENT role is not simply *"Prodi"* but the whole phrase.

**Special constructions.** Some linguistic constructions are particularly important for RTE, and have been treated by specific guidelines.

- *Support and copula verbs*. Supports and copulas (such as *be* and *seem*) are verbs carrying a minimal semantic content, which are used only to syntactically support a frame evoking noun. For example, in the sentence *"the President makes a statement"*, the verb *"makes"* supports the noun *"statement"*. We treat support and copula verbs as suggested in the FrameNet project annotation guidelines. We annotate the noun as FEE, leaving aside the verb (e.g., in the example above the word *"statement"* is used to evoke the frame STATEMENT). The same applies for copulas, as shown in Figure 4.

- *Existential construction*. Occurrences of the construct *"there be"* are annotated as FEE evoking the frame EXISTENCE only when the existential situation is the only meaning conveyed by the sentence, as in *"There are 11 official EU languages"*. This annotation guarantees that a minimum piece of semantic information (that of existence) is always conveyed by the annotation, allowing simple existential reasoning over a $(T, H)$ pair.

- *Modal expressions* (e.g. modal verbs or particles as *maybe* and *perhaps*) are annotated as FEEs evoking the *Likelihood* frame only when the modal meaning is the prevalent information conveyed in the sentence, as in *"Bush said the victory may not be possible"*. By using the general LIKELIHOOD frame, we aim at highlighting possible modal triggers in the $(T, H)$ pairs, so that an RTE system can easily spot them in texts and apply modal reasoning on the pair.

- *Metaphors*. In case of metaphors, it is possible to annotate with two different frames: a *source frame* to represent the literal meaning, and a *target frame* to represent the figurative meaning. We decided to annotate only with the target meaning, as this represents the real situation which is interesting for deriving the entailment. If there is no frame for the target meaning, we use the UNKNOWN frame.

## 4. FATE annotation

In this section we first describe the annotation process, and then present some statistics on the produced corpus.

### 4.1. Annotation process

The annotation has been carried out on the RTE-2 challenge test set (Bar-Haim et al., 2006), consisting of 800 $(T, H)$ pairs, 400 positive entailment examples and 400 negatives. Pairs are organized in 4 balanced subsets of 200 pairs built using different methods: information extraction (IE), question answering (QA), information retrieval (IR) and text summarization (SUM). The corpus accounts for a total of 28.684 word tokens.

We focused on the test set of RTE-2 for time constraints, leaving annotation of the corresponding development set and possibly other RTE datasets as a future work. However, according to different studies, the RTE-2 development and test set are quite similar and balanced in modeling different phenomena. Therefore, conclusion drawn on the test should by and large carry over to the development set.

We annotate frame-semantic information on top of the syntactic structure produced by the Collins parser, with a single flat tree for each frame. The root node is labeled by the frame name, the edges are labeled with the names of the frame elements. Annotation is performed using the SALTO graphic tool (Burchardt et al., 2006b). The tool displays the syntactic interpretation of texts, thus providing user-friendly functionalities to speed up the annotation. Frame and syntactic data are saved in SALSA/TIGER XML (Burchardt et al., 2006a).

Annotation has been done by an experienced annotator, initially trained and calibrated on a pilot dataset, supervised by a pool of expert researchers. To simulate the most natural annotation, texts and hypothesis have been shuffled and randomly reordered before the annotation. $T$ and $H$ of the same pair have been then annotated independently, i.e., the annotator is not influenced by its annotation on the $T$ when working on the $H$, and vice-versa. For resource and time constraints, we could not afford a full double-annotator process. Yet, we checked the consistency and the correctness of the final annotation, by implementing three different strategies.

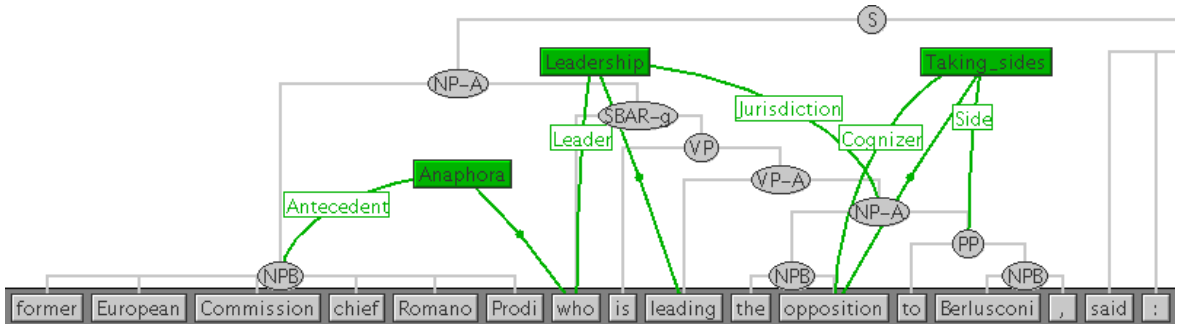


Figure 3: Example of *anaphora frame*.

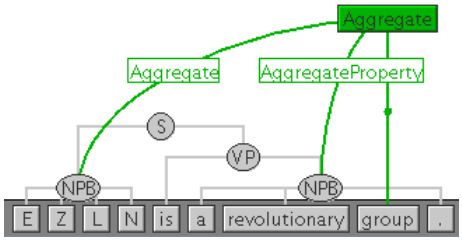


Figure 4: Example of treatment of a *copula construction*.

First, we performed an *inter-annotator agreement* test, asking a second experienced annotator to annotate 5% of the corpus (40 examples). We computed the agreement at three levels: FEE-agreement, frame-agreement and role-agreement. *FEE-agreement* is the percentage of commonly annotated FEE. *Frame-agreement* is the percentage of commonly selected frames, among those evoked by the same FEE by the annotators. *Role-agreement* is the percentage of commonly annotated roles (same name and same filler) among those belonging to commonly selected frames.[3] The obtained agreements are: 82% FEE-agreement, 88% frame-agreement, 91% role-agreement. These results indicate that the overall annotation is reliable. In particular, our definition of *relevant FEE* seems to be plausible and effective, as the two annotator selected the same FEEs in 82% of cases. Also, once the FEE has been selected, the tasks of finding the correct frame and the correct roles seems to be fairly easy and unambiguous. The sporadic cases of disagreement on frames usually involve the choice of different but highly similar frames (e.g. RISKY_SITUATION vs. RUN_RISK) or an unknown frame used by one annotator instead of the correct one present in the FrameNet hierarchy. Cases of disagreements on roles are generally due by one annotator missing a role.

As a second strategy to check consistency of the corpus, we computed an *intra-annotator agreement*. This has been made possible by the fact that the RTE-2 test dataset uses some sentences repeatedly across the dataset. We estimated the agreement over the positive corpus. In all, we counted 109 repetitions, i.e., pairs of repeated sentences. Over this set we computed a FEE-agreement of 97%, a frame-agreement of 98% and a role agreement of 88%, revealing a good level of consistency of the overall annotation.

Third, during the annotation process, we performed weekly *check meetings* between the annotator and the pool of supervisors, in which to report and discuss possible issues and inconsistencies.

### 4.2. Annotation statistics

The whole annotation was carried out in 230 hours: 90 hours for the positive examples, 140 hours for the negatives. In average, it took 13 minutes to annotate a positive pair, and 21 to annotate a negative. As positive and negative examples have in average the same number of tokens, these statistics clearly show the contribution of the span-annotation on speeding up the process.

In all, 4,489 frames were annotated, 1,666 in the positive set and 2,823 in the negative set. The average number of frames per pair is 5.6. The total number of roles is 9,518, namely 3,516 in the positive set and 6,002 in the negative set. The average number of roles per frame is 2.1. Table 2 reports the most frequent frames occurring in the corpus, and their number of occurrences. This list gives a general idea about the semantic domain characterizing the RTE-2 corpus, mostly referring to killing, disasters and competition events.

The annotation contains 373 *Unknown-frame* instances, accounting only for the 8% of the total frames. *Unknown roles* are 1% of the total roles. This means that FrameNet

[3] More in particular, to compute *FEE-agreement*, for each annotator we divide the number of FEE by the number FEE shared with the other annotator. Then, we compute the average. To compute *frame-agreement*, for each annotator we consider the frames which have been evoked by an FEE shared with the other annotator. Then, we compute the percentage of those frames that have been evoked also by the other annotator. Finally, we compute the percentage average between the two annotators. To compute *role-agreement* we consider only the roles belonging to frames in common between the annotators (same evoking FEE and same frame name). Then, we compute the percentage of these roles that have the same name and the same lexical fillers. Finally, we compute the percentage average between the two annotators.

| | | | |
|---|---|---|---|
| LEADERSHIP | 196 | ATTEMPT | 40 |
| STATEMENT | 152 | BEING_EMPLOYED | 38 |
| KILLING | 92 | CAUSATION | 36 |
| PEOPLE_BY_VOCATION | 90 | DEATH | 35 |
| CHANGE_POSITION_ON_A_SCALE | 85 | INTENTIONALLY_CREATE | 35 |
| ATTACK | 73 | BUSINESSES | 34 |
| FINISH_COMPETITION | 68 | EDUCATION_TEACHING | 33 |
| BEING_LOCATED | 51 | HOSTILE_ENCOUNTER | 31 |
| EVENT | 50 | PROTECTING | 31 |
| MILITARY | 49 | ACTIVITY_START | 29 |
| SURPASSING | 46 | BECOMING_AWARE | 29 |
| USING | 46 | MEANS | 29 |
| CAUSE_CHANGE_OF_POSITION_ON_A_SCALE | 45 | CAUSE_HARM | 28 |
| AGGREGATE | 43 | LOCALE_BY_USE | 26 |
| MEDICAL_CONDITIONS | 42 | BEHIND_THE_SCENES | 25 |

Table 2: Most frequent annotated frames in the RTE-2 test set.

coverage for the RTE corpus is surprisingly good. These numbers differ from figures reported, for example, for Salsa's German corpus annotation (Burchardt et al., 2006a), where one third of the verb occurrences could not be annotated with available FrameNet frames (largely due to the incompleteness of the frame inventory, not to cross-lingual differences). One possible reason for the discrepancy may be that only relevant frames have been annotated in FATE. Also, the annotators of FATE were allowed to annotate frames that looked appropriate in a rather flexible way, while the Salsa annotation for German followed a stricter annotation guideline. All in all, the 8% coverage lack of FrameNet frames indicates that the current FrameNet repository offers a good coverage over the RTE corpus. We can then conclude that coverage is unlikely to be a relevant issue limiting the application of FrameNet to RTE, as hypothesized in the Introduction.

The FATE corpus is available in XML SALSA/TIGER format at: `http://www.coli.uni-saarland.de/projects/salsa`.

## 5. Conclusions

In this paper we presented FATE, a manually frame-annotated corpus of textual entailment pairs, built over the RTE-2 challenge test set. To carry out the annotation, we introduced a novel FrameNet annotation schema, based on full-text annotation of so called *relevant FEEs*. The corpus offers a basis for addressing a number of unanswered research questions in the context of both using predicate argument structure for language processing and modeling textual inference.

As a first result, corpus statistics we provided show that FrameNet coverage is unlikely to be a major cause of the low performance so far obtained by inference systems based on frame semantics. As a future work, we aim at investigating other possible causes of FrameNet-based systems' low performance. First, we will leverage FATE to evaluate the accuracy of shallow semantic parsers over the RTE data. Moreover, we will closely inspect the corpus to look for regularities that characterize entailment at the frame level. In the medium term, we plan to model these regularities into an actual system for RTE, experimenting with both rule-based and machine learning approaches.

## Acknowledgments

Thanks to Konstantina Garoufi for providing the span annotation and to Alexander Fleisch for leading the annotation work. This work has partly been funded by the German Research Foundation DFG (grant PI 154/9-3).

## 6. 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, Idan Szpektor, and Oren Glickman. 2005. Definition and analysis of intermediate entailment levels. In *Definition and Analysis of Intermediate Entailment Levels*, Ann Arbor, Michigan.

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. 2005. Combining shallow and deep nlp methods for recognizing textual entailment. In *Pascal, Proceedings of the First Challenge Workshop, Recognizing Textual Entailment.*, Southampton, UK.

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.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal. 2006a. The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of LREC 2006*, Genoa, Italy.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal. 2006b. Salto – a versatile multi-level annotation tool. In *Proceedings of LREC 2006*, Genoa, Italy.

Aljoscha Burchardt, Nils Reiter, Stefan Thater, and Anette Frank. 2007. A semantic approach to textual entailment: System evaluation and task analysis. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague.

M. Collins. 1999. *Head-driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence D'Alché-Buc, editors, *Evaluating Predictive Uncertainty, Visual Object Categorization and Textual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pages 1–27, Heidelberg, Germany. Springer.

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

K. Garoufi. 2007. Towards a better understanding of applied textual entailment: Annotation and evaluation of the rte-2 dataset. *M.Sc. thesis*, Saarland University.

Paul Kingsbury, Martha Palmer, and Mitch Marcus. 2002. Adding semantic annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference*, San Diego.

Ken Litkowski. 2006. Componential analysis for recognizing textual entailment. In *Proceedings of PASCAL RTE2 Workshop*.

S. Pado. 2007. *Cross-Lingual Annotation Projection Models for Role-Semantic Information*. Ph.D. thesis, Saarland University, Germany.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, and Christopher R. Johnson. 2007. Framenet: Theory and practice. Available at `http://framenet.icsi.berkeley.edu/`.

Satoshi Sekine, Kentaro Inui, Ido Dagan, Bill Dolan, Danilo Giampiccolo, and Bernardo Magnini, editors. 2007. *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Association for Computational Linguistics, Prague, June.