Extending Fine-Grained Semantic Relation Classification to Presupposition Relations between Verbs

Galina Tremper and Anette Frank
Department of Computational Linguistics
Heidelberg University, Germany

Abstract

In contrast to typical semantic relations between verbs, such as antonymy, synonymy or hyponymy, presupposition is a lexical relation that is not very well covered in existing lexical resources. It is also understudied in the field of corpus-based methods of learning semantic relations. But presupposition is very important for the quality of automatic semantic and discourse analysis tasks. In this paper we present a corpus-based method for learning presupposition relations between verbs, embedded in a discriminative classification approach for fine-grained semantic relations. The focus of the present paper is to discuss methodological aspects of our approach including the choice of resources and data sets, the selection of features for classification, and design decisions regarding the annotation of fine-grained semantic relations between verbs.

1 Introduction

Determining lexical-semantic and discourse-level information is crucial in event-based semantic processing tasks. This is not trivial, because significant portions of content conveyed in a discourse may not be overtly realized. Consider the examples (1) and (2), where (1) bears a presupposition that is overtly expressed in (2):

(1) *Spain won the finals of the 2010 World Cup.*

(2) *Spain played the finals of the 2010 World Cup.*

The presupposition expressed in (2) is implicitly encoded in (1), through lexical knowledge about the verb *win*, and is thus automatically understood by humans who interpret (1), given their linguistic knowledge about the verbs *win* and *play*.

One reason for addressing presupposition detection as a discriminative classification task is that *presupposition* needs to be carefully distinguished from other lexical relations, in particular *entailment* - as the two relations are closely related, but crucially differ in specific aspects. Consider the sentence pair (3) and (4).

(3) *The president John F. Kennedy was assassinated.*

(4) *The president John F. Kennedy died.*

Sentence (3) logically entails (4). But how does this differ from the presuppositional relation between (1) and (2)?

The differences between *presupposition* and *entailment* can be studied using special presupposition tests (Levinson, 1983). The most compelling one, which we will use throughout, is the negation test. It shows that the presupposition relation is preserved under negation, while entailment is not. Applied to (1) and (3), we note that (5), the negation of (1), still implies (2), while (6), the negation of (3), does not imply (4):

(5) *Spain didn't win the finals of the 2010 World Cup.*

(6) *The president John F. Kennedy was not assassinated.*

This can be taken as evidence that *win* presupposes *play*, while *assassinate* logically entails *die*. Thus, the negation test not only helps us to distinguish these closely related verb relations, it also points to the crucially distinct logical behaviour of these relations in deriving implicit meaning from discourse, which is the main motivation underlying our work.

Similar to entailment, presuppositional relations between verbs are essentially grounded in world knowledge. At the same time, they are crucial for the computation of discourse meaning and inference, and thus, need to be captured in large-scale lexicons, along with more structural, taxonomic semantic relations, such as *antonymy*, *synonymy*, or *hyponymy*. The latter are the primary relations that make up the WordNet database (Fellbaum, 1998). By contrast, *entailment*, *presupposition* and other more fine-grained relations are not covered in sufficient detail. Chklovski and Pantel (2004) were first to attempt the automatic classification of fine-grained verb semantic relations, such as *similarity*, *strength*[1], *antonymy*, *enablement*[2] and *happens-before* in VerbOcean. In the present paper we aim to extend the classification of semantic relations between verbs to *presupposition*. To our knowledge, it has not been attempted before. We will address this task in a discriminative classification task – by distinguishing presupposition from other semantic relations, in particular *entailment*, *temporal inclusion* and *antonymy*.

Our overall aim is to capture implicit lexical meanings conveyed by verbs, and to use this knowledge by making it explicit for improved discourse interpretation. This overall aim can be divided into two tasks:

Detecting and discriminating fine-grained semantic relations: We first detect and distinguish between fine-grained semantic relations including presupposition, at the type level, to encode this lexical knowledge in lexical semantic resources.

Deriving implicit meaning from text: In a second step, we will apply this knowledge for the interpretation of discourse, at the token level, in order to enrich the overtly expressed content with implied, implicit knowledge, conveyed by presupposition, entailment, or other lexical semantic relations. This kind of information can contribute to improving the quality of automatic semantic and discourse processing tasks, such as information extraction, text summarizarion, question-answering and full-fledged textual inferencing or natural language understanding tasks.

In the present paper we concentrate on the first task. We present a corpus-based method for learning semantic relations between verbs with a special focus on presupposition. The structure of the paper is as follows: Section 2 reviews related work.

---

[1]*strength*: V_1 are V_2 similar, but V_1 denotes a more intense action (Chklovski and Pantel, 2004)

[2]*enablement*: V_1 makes V_2 possible (Barker and Szpakowicz, 1995)

Section 3 studies the space and discriminative properties of fine-grained semantic relations, and introduces the basic method and selected features for classification and annotation strategies. In Section 4, we report our classification experiments. We introduce the resources used and discuss different annotation strategies. We present two corresponding classification experiments and the results we obtain. Section 5 offers a detailed error analysis regarding the used resources, features and annotation design schemes. Finally, we summarise and present objectives for future work in Section 6.

2 Related Work

Significant progress has been made during the last decade in the automatic acquisition of semantic relations between verbs using corpus-based methods. Lin and Pantel (2001) proposed a distributional method for extracting highly associated verbs. This method retrieves verb pairs which are linked by a semantic relation, but does not identify the type of these semantic relations. Their work was used as a starting point to automatically classify fine-grained semantic relations in other projects, such as VerbOcean (Chklovski and Pantel, 2004). Chklovski and Pantel (2004) used a semi-automatic pattern-based approach for extracting fine-grained semantic relations between verbs, including *similarity*, *strength*, *antonymy*, *enablement* and *happens-before*.

In a related strand of work, many projects tried to generate textual entailment rules (e.g. Pekar (2008), Ben Aharon et al. (2010)), however, they do not subclassify the extracted entailment pairs in *presupposition*, *entailment*, *cause* or other classes. Berant et al. (2010) try to improve on learning isolated textual entailment rules.

Only little work is devoted to the computational treatment of presupposition. Bos (2003) adopted the algorithm of van der Sandt (1992) for presupposition resolution. His approach is embedded in the framework of DRT (Kamp and Reyle, 1993). It requires heavy preprocessing and a lexical repository of presuppositional relations. Clausen and Manning (2009) compute presuppositions in a shallow inference framework called "natural logic". Their account is restricted to computing factivity presuppositions of sentence embedding verbs. In the field of corpus-based learning of semantic relations, the automatic acquisition of presupposition relations remains understudied.

3 A Corpus-based Method for Learning Semantic Relations

We present a corpus-based method for learning semantic relations between verbs including the presupposition relation. We subclassify verb pairs into five classes of relations: *presupposition*, *entailment*, *temporal inclusion*, *antonymy* and *other/unrelated*. The verbs in the last class stand in no or some semantic relation not considered here. In a preliminary step, we also considered the *synonymy* relation.

For classification, we start with a small number of seed verb pairs selected manually for each semantic relation and used to build a labeled corpus for training of binary

131

feature-based classifiers, one for each semantic relation. These classifiers are applied to a large set of unlabeled verb pairs. The candidate verb pairs are selected from a set of semantically related verbs according to the DIRT collection (Lin and Pantel, 2001). For the chosen candidate verb pairs, we extract corpus samples for feature-based classification in which both verbs co-occur, using the ukWaC corpus (Baroni et al., 2009). At this step we excluded the synonymy relation, as synonymous verbs usually do not occur contiguously in a single sentence.[3] For the remaining five semantic relations, we independently train five binary classifiers, using the J48 decision tree algorithm (Witten and Frank, 2005). Each of the five classifiers is applied to each sentence from the unlabeled corpus. The predictions of the classifiers are combined using ensemble learning techniques to determine the most confident classification.

3.1 Properties of Semantic Relations between Verbs and Feature Set

In order to establish an effective feature set for the classification we analyzed the properties of the relations between the verbs we aim to distinguish: *presupposition*, *entailment*, *temporal inclusion*, *antonymy* and *synonymy*[4]. We observe that the paradigmatic lexical semantic relations like *antonymy*, *synonymy* and *temporal inclusion* typically do not involve a temporal order. In contrast, *presupposition* relations between verbs involve a temporal sequence. The event that is presupposed tends to precede the event that triggers the presupposition. The verbs which stand in an *entailment* relation may or may not involve a temporal order; in case of temporal sequence the overtly realized verb can precede or succeed the entailed verb.

Another important aspect is the behaviour of the different semantic relations under negation. Some semantic relations (e.g. *presupposition* and *temporal inclusion*) are preserved under negation. In this way they can be distinguished from other semantic relations (e.g. *entailment* or *synonymy*) which do not persist under negation.

		Behaviour under Negation			
		$V_1 \to V_2$	$\neg V_1 \to V_2$	$V_1 \to \neg V_2$	$\neg V_1 \to \neg V_2$
Temporal Sequence	V_1 precedes V_2	E			E
	V_1 succeeds V_2	P E	P		P E
	No temporal sequence	E T S	T A	A	E T S

Table 1: Properties of the Semantic Relations:
P(resupposition), E(ntailment), T(emporal Inclusion), A(ntonymy), S(ynonymy)

---

[3]An analysis of sentences in which synonymy were found to co-occur shows that the verbs appear only accidentally within a single sentence, and should therefore be classified as unrelated. We therefore eliminated synonymy from the set of target relations.

[4]While we will classify synonyms as unrelated in our experiments, for completeness we do include synonymy in this discriminative analysis.

The distinguishing temporal and negation properties that cross-classify these semantic relations are schematically represented in Table 1[5]. As shown in Table 1, it is possible to distinguish the targeted semantic relations on the basis of these properties:

(i) *Presupposition* and *entailment* (whether or not temporally related) are distinguished on the basis of persistence under negation, which holds for *presupposition* only. The same pattern holds for *temporal inclusion* vs. *entailment*.

(ii) *Temporal inclusion* and *presupposition* behave alike regarding negation properties, but can be distinguished in terms of temporal sequencing properties.

(iii) *Synonymy* and *entailment* are difficult to distinguish in cases where *entailment* does not involve temporal sequence. However, since we exclude *synonymy* and range it under the class *unrelated*, this does not cause a problem.

(iv) *Antonymy* behaves clearly different from *entailment* and *presupposition* wrt. both properties, and from *temporal inclusion*, regarding negation properties.

(v) For completeness, *antonymy* and *synonymy* are opposites to each other wrt. negation properties, if we considered *synonymy* as a target relation.

Thus, the properties pointed out above could be used to distinguish the four target semantic classes. These four classes, in turn, need to be distinguished from the fifth class of unrelated verb pairs - which will include synonymous verbs, in case they (accidentally) co-occur. That is, we will need to model contextual relatedness features, to distinguish between the target relation classes and the class of unrelated verb pairs, and accidentally co-occuring verb pairs. For this purpose we will propose rather abstract contextual boundedness features that are able to characterize a broad variety of constructions that may be indicative for (any of) the targeted relation classes. We will refer to these features as "contiguity features".

3.2 Features for Classification

Temporal Sequence. To detect the distinct temporal relations between verbs we made use of features similar to the feature set used by Chambers et al. (2007):

1. Verb form (tense, aspect, modality, voice, negation, etc.).
2. PoS contexts (two words preceding and two words following each verb).

Further features we used for determining temporal relations are:

3. Coordinating/subordinating conjunctions.
4. Adverbial adjuncts.

Negation. Our analysis of the properties of the semantic relations shows that negation is crucial for distinguishing our target relations. Currently, we use as a trigger for

---

[5]Example of using the table: V_1 is a placeholder for the trigger verb and V_2 — for the inferred verb. For the *presupposition* verb pair *(win,play)*, the event of winning sth (V_1) typically temporally succeeds the event of playing something (V_2), therefore we concentrate on the second row. The event of winning something implies the event of playing something ($V_1 \rightarrow V_2$). The event of not winning something could be interpreted in two ways: constancy under negation — not winning although playing ($\neg V_1 \rightarrow V_2$) or cancellation — not winning because of not playing ($\neg V_1 \rightarrow \neg V_2$).

the negation feature the presence or absence of the negative particle *not/n't* (as part of the verb complex). In future work we plan to integrate further negation properties such as negation adverbs or suffixes.

Contiguity. One important task in the subclassification of verb relations is to decide whether or not two verbs stand in one of the targeted meaning relations in a given context. We observed that besides the distance between the verbs, the co-referential binding of the verb arguments can be very useful in determining contextual contiguity of verb pairs in specific contexts. Finally, in case of ambiguous verb readings, subcategorisation frames help to restrict a given verb relation to specific verb meanings. The following features were investigated for this purpose:

1. The distance between two analyzed verbs and the order of their appearance.
2. The number of main verbs occurring between two analyzed verbs.
3. The length of the path of grammatical functions relating the two verbs.
4. Co-reference relation holding between the subjects and objects of the verbs (both verbs have the same subject/object, the subject of one verb corresponds to the object of the second or there is no relation between them).
5. Subcategorization frames for two analyzed verbs.

4 Experiments and Results

4.1 Resources

In our experiments, we made use of the following resources.

1. *ukWaC* is the English part of the WaCKy corpora (Baroni et al., 2009). The corpus was constructed by crawling the .uk Internet domain and contains more than 2 billion tokens. Currently, it is the largest freely available resource for English that includes PoS and lemmatisation information. We use this corpus for extracting the training and test data sets, because it is large enough for obtaining high precision corpus data using statistical methods. ukWaC is certainly smaller than the entire English Web, but given that it is enriched with PoS and lemma annotations, multiple Internet queries can be replaced by a single one that applies to the pre-analysed ukWaC corpus. In our experiments we used the first three parts of the ukWaC corpus which contain about 280 million sentences.
2. Taking into account all possible combinations of verbs acquired from *ukWaC* yields an extremely large set of candidate pairs for classification, and the amount of unrelated verbs pairs would be huge. Instead, we used the DIRT-Collection to select pairs of highly associated verbs as candidates for classification. The *DIRT-Collection* (Lin and Pantel, 2001) is the output of the paraphrasing algorithm called DIRT applied on 1 GB of newspaper text from the TREC collection. It consists of pairs of verbs that have been determined to stand in a semantic relation using corpus-based association measures. DIRT contains 5,604 verb types and 808,764

verb pair types. We filtered the verb pairs extracted from DIRT using a threshold applied on the verb pair frequencies of appearance[6] and applied the PMI test with threshold 2.0. This reduces the number of candidate verb pairs to 199,393.

4.2 Annotation strategies for establishing a Gold Standard

Annotating semantic relations, especially implicit relations like *presupposition* and *entailment* is a difficult task because of the subtlety of the tests and the involved decisions. In order to obtain reliable annotations it is important to define the task in an easy and accessible way and to give clear instructions to the annotators.

We decided to formulate two annotation tasks: one on the level of verb pairs given as types out of context (type-based annotation) and another on the level of verb pairs in context (token-based annotation) and to examine to what degree the results obtained from the two annotation setups correlate.

4.2.1 Gold Standard 1 (GS1): Type-based annotation

The complete set of verb pair candidates (about 200,000 verb pairs) is impossible to annotate manually, therefore we randomly selected a small sample of 100 verb pairs. In order not to influence the judges' decisions, we eliminated the system annotations. Since some verbs can have more than one meaning and consequently verbs in a verb pair can stand in more than one semantic relation, the judges were allowed to assign more than one relation to each verb pair.

To support the annotators in their decisions, we provided them with a couple of inference patterns and examples for each semantic relation. This is shown in Table 2.

Sem. Relation	Pattern	Example	*Substitution in pattern*
Presupposition	V_1 presupposes V_2, *not* V_1 presupposes V_2	*win - play*	*winning* presupposes *playing* *not winning* presupposes *playing*
Entailment	V_1 implies V_2, *not* V_1 doesn't imply V_2	*kill - die*	*killing* implies *dying* *not killing* doesn't imply *dying*
Temporal Inclusion	V_1 happens during V_2 or V_1 is a special form of V_2	*snore - sleep* *mutter - talk*	*snoring* happens during *sleeping* *muttering* is a special form of *talking*
Antonymy	either V_1 or V_2, V_1 is the opposite of V_2	*go - stay*	either *going* or *staying* *going* is the opposite of *staying*
Other/unrelated	none of the above	*jump - sing*	

Table 2: Semantic Relations and Inference Patterns for Annotation

The inter-annotator agreement for this task was 63% corresponding to a Kappa value of 0.47. This can be taken as an indication for a high difficulty of semantic relation

---

[6]The verb pair frequencies were calculated only for the first three parts of the ukWaC corpus.

annotation when performed out of context.

4.2.2 Gold Standard 2 (GS2): Token-based annotation

Since type-based annotation turned out to be very difficult, we decided to simplify the task by providing the annotators with verb pairs in their original context. For this token-based annotation we took the same 100 verb pairs and randomly selected 5 to 10 contexts for each of them (the total number of all contexts was equal to 877). Similar to the type-based annotation task we eliminated all system labels. In contrast to type-based annotation, we only accepted a single relation label for a given verb pair.

The inter-annotator agreement for this task was 77.4%, which corresponds to a Kappa value of 0.44. Error analysis showed that the most important problems are not due to semantic relations which are difficult to distinguish (e.g. *presupposition* and *entailment*), but rather in determining whether or not there is a specific semantic relation between two verbs in a given context.

We examined the correlation between the type- and token-based Gold Standards by comparing the annotations of a single judge for both annotation tasks. For 62% of verb pair types we observe an overlap of labels, 28% of the verb pair types were assigned labels on the basis of the annotations in context which were not present on the type level without context, or the type level label was not assigned in context, because of the small amount of contexts for a verb pair. For 10% of verb pair types we found conflicting annotations (for example, *presupposition* and *entailment*). Thus, for the most part (62%) the type-based annotation conforms with the ground truth obtained from token-based annotation. An additional 28% of verb pairs can be considered to be potentially correct. The divergences for these verb pairs could be explained by the random procedure of the context extraction which does not always return appropriate contexts. They can also be explained by the difficulty for the annotator to consider all possible verb meanings for highly ambiguous verbs in type-based annotation.

4.2.3 Gold Standard 3 (GS3): Type-based annotation deduced from GS2

Since our ultimate goal is to detect and distinguish fine-grained semantic relations at the type level, we used the token-based annotations to deduce type-based annotations. For GS1 we accepted multiple relation labels. Therefore, for constructing GS3, for each verb pair type we selected up to three most probable annotations (most frequent annotations from the token-based annotations of GS2). An exception was made for the *other/unrelated* class: only the verb pairs annotated unambiguously in all cases with the *other/unrelated* label were considered to belong to this class.

The distribution of semantic relations in all three Gold Standards is given in Table 3.

Semantic Relation	Frequency in GS1	Frequency in GS2	Frequency in GS3
Presupposition	18	70	24
Entailment	8	44	8
Temporal Inclusion	19	26	12
Antonymy	12	44	10
Other/unrelated	43	693	46

Table 3: Distribution of Semantic Relations in Gold Standards (GS)

The distribution of relation types in GS1 and GS3 is very close. Because GS3 was derived from GS2 by selecting up to three most probable annotations, the overlap between them is identical to the overlap between GS1 and GS2 discussed above (62%)[7]. A confusion analysis shows that the set of verb pairs labeled as *entailment* remains stable (*entailment* and *presupposition* are confounded in only two cases). Annotation in context reduces the number of *temporal inclusion* and *antonymy* relations that were annotated out of context. On other hand, we observe a tendency to annotate more verb pairs with the *presupposition* relation.

4.3 Best Features for Classification

Our final classification is based on five binary classifiers, one for each semantic relation. We analyzed which of the features from the feature set (see Section 3.2) are the most effective for determining each semantic relation. We also compared the best features for binary classifiers with the best features for single multi-class classification. The best features were determined on the basis of the manually annotated training set using Gain Ratio coefficient. The top five best performing features for each individual semantic relation and for the full set of relations are presented in Table 4.

Table 4 shows that conjunctions between verbs are important for all semantic relations. For determining presupposition, the verb that triggers the presupposition (V_1) seems to be more important than the presupposed verb (V_2). By contrast, for determining the entailment relation, the verb which is the logical consequence (V_2) seems to be more important than the verb which implies it (V_1). The selected features highlight the importance of coreference relations holding between arguments, as well as the subcategorization frame information for detecting a specific semantic relation between verbs. They characterize in particular the unrelated class, and *antonymy*, as contextually unrelated verbs[8]. Negation was not selected as a strong feature, although it prominently figures in our analytical cross-classification scheme. This may be due to sparseness,

---

[7]For computing overlap we consider all relations annotated per type.

[8]This suggests exploring a two stage classification that in a first step distinguishes unrelated verbs from related ones, and subsequently classifies the remaining fine-grained semantic relations.

Semantic Relation	Top-5 Best Features
Presupposition	Order, Conj, AdvAdj of V_1, Mod of V_1, SubCat of V_1
Entailment	Order, Conj, AdvAdj of V_2, Mod of V_2, Asp of V_2
Temporal Inclusion	Conj, AdvAdj of V_1, SubCat of V_1, SubCat of V_2, Dist
Antonymy	Conj, AdvAdj of V_2, SubjObj, NumVerb, Dist
Other/no	Conj, SubjObj, SubCat of V_1, SubCat of V_2, GF-Length
All	Order, Conj, AdvAdj of V_1, SubCat of V_1, SubjObj

Table 4: Top-5 Best Features
V_1, V_2 – placeholders for verbs in the verb pair, Order – Order of appearence, Conj – Conjunction, AdvAdj – Adverbial adjunct, Mod – Modality, Asp – Aspect, Dist – Distance between verbs, GF-Length – length of GF-path between verbs, NumVerb – number of intervening main verbs, SubCat – Subcat frame, SubObj – Coreference between Subject/Object

given the restricted feature set currently used for characterizing negation properties.

4.4 Classification

Starting with a small number of seed verb pairs (3 to 6) (see beginning of Section 3), we build a training corpus consisting of three parts: a manually annotated training set (5,032 sentences) collected from the ukWaC for the seed verb pairs, a heuristically annotated extended training set (9,918 sentences)[9] and heuristically annotated synonymous verb pairs in context (757 sentences)[10]. The set of unlabeled verb pairs in context is built from the filtered set of related verb pairs from DIRT (see Section 4.1), and includes about 4,500,000 sentences. For the classification we use the outputs of five binary J48-classifiers independently applied on the same set of unlabeled data.[11]

4.5 Experiments and Results

We performed two experiments for the classification of verb pairs. In the first experiment we classified each candidate verb pair in context (token-based classification) and evaluated the results against GS2. In the second experiment we classified all candidate verb pairs at the type level, by aggregation from instance-level classifications in context (type-based classification) and evaluated the results against GS1 and GS3.

---

[9]Heuristic annotation was performed using a manually compiled small stoplist of patterns meant to eliminate wrong instances (see Tremper (2010) for details). In future work we will explore the use of classifiers trained on shallow features (Banko et al., 2007).

[10]The synonyms were obtained from WordNet (Fellbaum, 1998).

[11]We also experimented with classification using a multiclass J48-classifier. Due to the lower results on a small subset of the manually annotated training corpus, we didn't evaluate this classifier on the unlabeled data set.

4.5.1 Experiment 1

To perform **token-based classification** we determined the most confident classification for each instance of the unlabeled verb pair in context using a voting architecture. We compared the classifications of all five binary classifiers and selected the classification with the highest confidence.[12]

We evaluated the results against the token-based Gold Standard 2 (see Section 4.2.2). We computed precision, recall and f-score. As baseline we took the distribution found in the manually labeled gold standard as the underlying verb relation distribution. The results for each semantic relation are shown in Table 5.

Semantic relation	Precision	Recall	F-Score	Baseline
Presupposition	23%	27%	25%	8%
Entailment	18%	25%	21%	5%
Temp. Inclusion	10%	12%	11%	3%
Antonymy	42%	68%	52%	5%
Other/Unrelated	73%	59%	65%	79%
Macro-Average	33%	36%	34%	
Micro-Average	59%	54%	56%	

Table 5: Evaluation of the Results of Experiment 1

Except for the unrelated class, the results are well above the baseline. The results show that typical and broad semantic relations such as *antonymy* perform better than *presupposition* and *entailment*. *Temporal inclusion* achieves the lowest results for token-based classification. Here, the error analysis shows that this relation was most often confounded with unrelated verb pairs. Some examples of the correct and wrong classifications are presented in the Table 6.

4.5.2 Experiment 2

To perform **type-based classification** we first performed token-based classification as described in Experiment 1. We combined the results obtained for individual instances to derive relation labels on the type level as follows. We eliminated semantic relation labels which were assigned to less than 10% of the instances of a given verb pair. Verb pairs which after this step were assigned more than three semantic relations are ignored (remain unclassified). Finally, verb pairs that were left with up to three semantic relations, each of which was assigned to at least 10% of the examples, were labeled with all of these semantic relations.

We evaluated the results against the type-based Gold Standard 1 (see Section 4.2.1) and Gold Standard 3 (see Section 4.2.3). Again we report precision, recall and f-score.

---

[12]Only the classifications with a confidence exceeding 0.75 were accepted for voting.

Sem. Relation	Verb pair	Correct classification	Wrong classification (System label)
Presupposition	*classify – identify*	It was noted that of the thirteen issues identified in the report eight were classified as high priority.	The meeting focussed on issues of identifying, classifying and marking up names in both corpora and analytical projects. (Temp. Inclusion)
Entailment	*click – send*	Clicking the Send feedback button will send any feedback you have entered.	You can send us your comments by simply clicking on this email. (None)
Temp. Inclusion	*reply – say*	Replying to the toast to the guests, Dr Julia King said how privileged the Faculty was to have two such active alumni associations.	18 out of the 20 Rehabilitation Officers who replied said that there is somewhere they can take clients for equipment demonstrations. (None)
Antonymy	*disconnect – connect*	A click should be heard every time the antenna wire is connected or disconnected.	This allows you to connect and disconnect easily , simply by clicking on the icon and selecting the relevant option. (None)

Table 6: Examples of the correct and wrong classifications in context

In contrast to token-based classification, we considered verb pairs to be correctly labeled if at least one tag was correct. The results are shown in Table 7.

Semantic relation	Gold Standard 1				Gold Standard 3			
	Prec.	Recall	F-Score	Baseline	Prec.	Recall	F-Score	Baseline
Presupposition	43%	33%	37%	18%	50%	29%	37%	24%
Entailment	36%	50%	42%	8%	36%	50%	42%	8%
Temp. Inclusion	50%	16%	24%	19%	33%	17%	22%	12%
Antonymy	75%	75%	75%	12%	58%	70%	63%	10%
Other/Unrelated	56%	74%	64%	43%	68%	85%	76%	46%
Macro-Average	33%	50%	40%		49%	50%	49%	
Micro-Average	53%	53%	53%		59%	59%	59%	

Table 7: Evaluation results for Experiment 2 (against Gold Standards 1 and 3)

The type-based classification clearly outperforms token-based classification. One of the reasons for the better performance of type-based classification is certainly that more examples are considered for assigning a relation, in which case voting plays a major role in eliminating unsecure decisions. By contrast, in token-based classification, each example is considered and labeled in isolation, including those with small confidence scores. The results for type-based classification are clearly exceeding the baseline for all relation types. Comparing evaluation against GS1 and GS3, the results for GS3 are higher, with a balanced macro-average in precision, recall and f-scores of about

140

50%, with clear improvement of precision for *presupposition*, a drop in performance for *antonymy*, and high performance gains for distinguishing the *unrelated* class.

5 Error Analysis and Discussion

5.1 Resources

Using the verb pairs extracted from the DIRT collection (see Section 4.1.), we extracted corpus samples from the ukWaC corpus (both for establishing labeled training and un-labeled test corpora). The PoS and lemma information encoded in ukWaC saves time needed to tag and lemmatise the corpus. But it also incurs errors that cause problems in the classification. An error analysis conducted on a small subset of the manually anno-tated training corpus shows that 10% of all errors are caused by erroneously annotating nouns or adjectives as verbs. This problem can be solved by using information from a deep parser to double check the PoS-tags provided by ukWaC.

5.2 Annotation

Comparing the results of the two annotation setups clearly shows that both are difficult, yet in different ways. Annotation on the type level is difficult because no indication is given about the intended meaning of the verbs. Hence the annotators need to consider all possible combinations of meanings for any pair of verbs. However, embedding the pairs in their original context doesn't make the decision much easier. This is because some sentences involve complex structure and interpretation difficulties, which require a lot of attention and time to annotate the individual examples.

To render the annotation task more reliable and less time-consuming, we need to develop an annotation strategy which includes the positive elements of both annotation strategies described above. One solution could be to present verb pairs with prototypi-cal arguments instead of taking the concrete sentence as a disambiguating context. The argument abstractions could be represented using WordNet hypernyms.

Another strategy could be to use a question scenario to collect annotations. The idea is to guide the annotator to verifying the discriminative categorizing properties "temporal sequence" and "persistence under negation", using a "setting" and a follow-up question that is intended to elicit the critical/missing piece of information needed to classify the verb pair in question. Using the properties of the semantic relations displayed in Table 1 we established a set of questions that elicit only three possible answers (Yes/No/Maybe). The answers can be used to distinguish between the target semantic relations and thus to annotate the data. (7)-(9) list examples of such questions:

(7) *X found Y. Did X search Y?*[13]

The answer *yes* in (7) excludes the semantic relation *antonymy* for the pair *find* and *search* (as antonyms can't be valid at the same time).

---

[13]**X** and **Y** in the questions are placeholders for arguments which can be refined using prototypical nouns.

(8)   *X didn't find Y. Did X search Y?*

The answer *maybe* in (8) indicates persistence under negation, and thus excludes the relation *entailment*.

(9)   *Did X find Y after searching?*

The answer *yes* in (9) excludes the relation *temporal inclusion* between *find* and *search*. On the basis of these three answers we can annotate the verb pair with *presupposition*. By exploiting the properties of the target relations regarding temporal sequence and negation, as summarized in Table 1, we can distinguish each of the 5 target classes with maximally three questions per verb pair. The decision tree for distinguishing between semantic relations is presented in Figure 1.
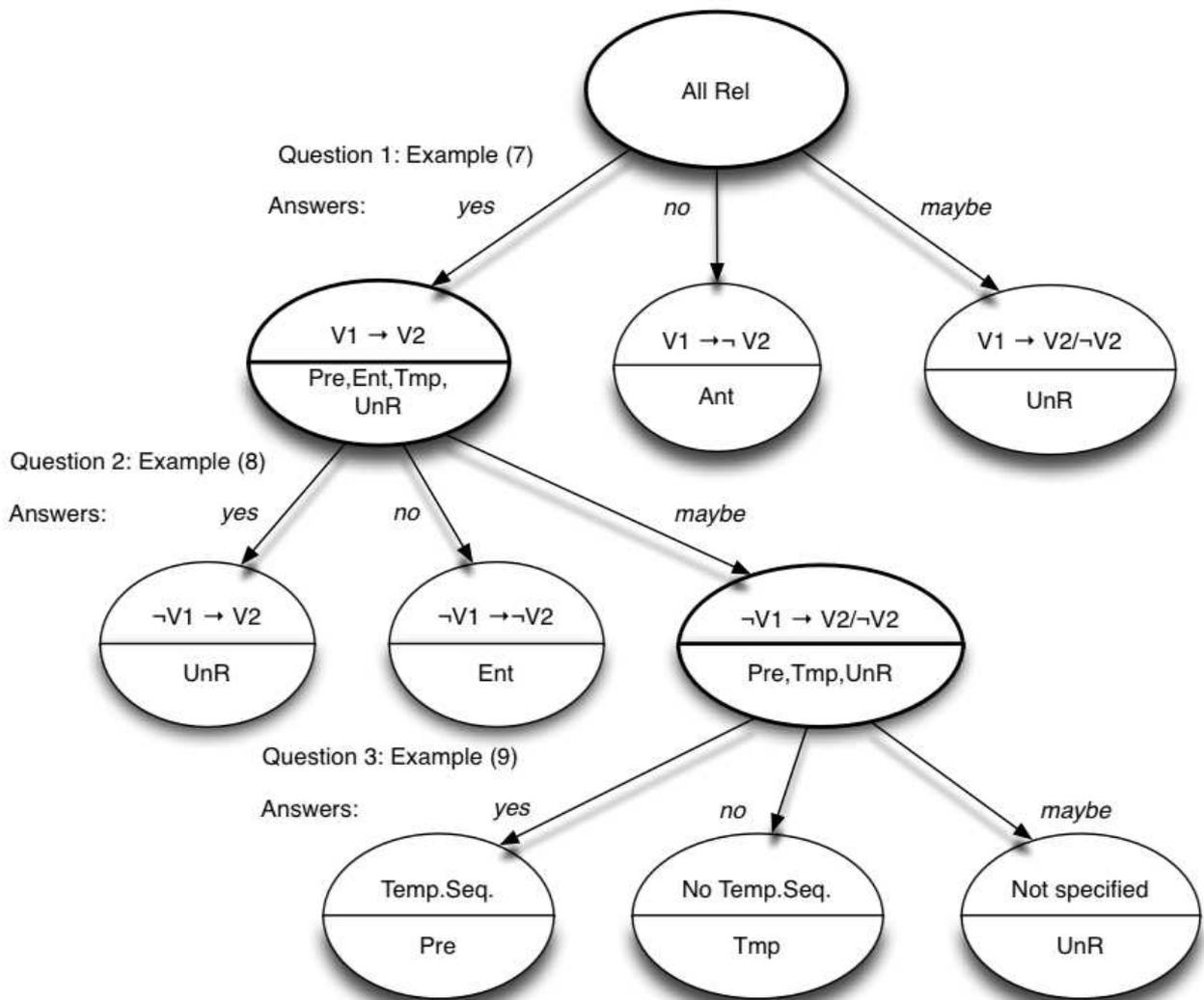

Figure 1: Decision Tree for distinguishing between semantic relations
Pre – presupposition, Ent – entailment, Tmp – temporal inclusion, Ant – antonymy, UnR – unrelated, Temp. Seq. – temporal sequence, V_1 & V_2 – placeholders for verbs

6   Conclusion and Future Work

In this paper we present first results in the corpus-based acquisition of presupposition relations between verbs, embedded in a discriminative classification approach for fine-grained semantic relation classification. We observe that presupposition is more difficult to determine than typical semantic relations like antonymy.

There are still many open issues left for future work. Coming up with solutions for word sense disambiguation and coreference resolution could help to eliminate the major source of observed errors. To improve the reliability of annotation and system performance, we plan to integrate information about predicate-argument structure using information extracted from FrameNet (Ruppenhofer et al., 2005) and VerbNet (Kipper, 2005) as well as prototypical argument head nouns encoding selectional preferences. We aim to improve classification performance by extending our feature set for characterizing negation properties. We also plan to evaluate the question-based annotation scenario proposed in Section 5.2. Given that it relieves the annotator from considering complex logical decisions, it could be appropriate for a crowd-sourcing annotation setup. We will also investigate a cascaded classification approach that follows the structure of the annotation decision tree.

The focus of the present paper was to describe in detail the underlying properties of the selected relations, our choice of resources and features for context-based classification, and to discuss design issues of the annotation task. Future work will establish an annotation and evaluation setup for the induction of implicit information in context, using the acquired semantic relations, in particular the presupposition relation pairs.

References

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open information extraction from the web. In *20th International Joint Conference on Artificial Intelligence*, pages 2670–2676, Hyderabad, India, 2007.

Ken Barker and Stan Szpakowicz. Interactive semantic analysis of clause-level relationships. In *Proceedings of PCLING 95*, pages 22–30, Brisbane, Australia, 1995.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, 43(3):209–226, 2009.

Roni Ben Aharon, Idan Szpektor, and Ido Dagan. Generating entailment rules from FrameNet. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 241–246, Uppsala, Sweden, 2010.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. Global learning of focused entailment graphs. In *Proceedings of the ACL 2010 Conference*, pages 1220–1229, Uppsala, Sweden, 2010.

Johan Bos. Implementing the binding and accommodation theory for anaphora resolution and presupposition projection. *Computational Linguistics*, 29(2):179–210, 2003.

Nathanael Chambers, Shan Wang, and Dan Jurafsky. Classifying temporal relations between events. In *Proceedings of the ACL-07 conference*, pages 174–176, Prague, Czech Republic, 2007.

Timothy Chklovski and Patrick Pantel. VerbOcean: Mining the web for fine-grained semantic verb relations. In *Proceedings of EMNLP 2004*, pages 33–40, Barcelona, Spain, 2004.

David R. Clausen and Christopher D. Manning. Presupposed content and entailments in natural language inference. In *Proceedings of the 2009 Workshop on Applied Textual Inference, ACL-IJCNLP 2009*, pages 70–73, 2009.

Christian Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, first edition, 1998.

Hans Kamp and Uwe Reyle. *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht, 1993.

Karen Kipper. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, 2005.

Stephen C. Levinson. *Pragmatics*. Cambridge: Cambridge University Press, 1983.

Dekang Lin and Patrick Pantel. Discovery of inference rules for question answering. *Natural Language Engineering*, 7:343–360, 2001.

Viktor Pekar. Discovery of event entailment knowledge from text corpora. *Computer Speech & Language*, 22(1):1–16, 2008.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. FrameNet II: Extended theory and practice. Technical report, ICSI, 2005. URL `http://framenet.icsi.berkeley.edu/book/book.pdf`.

Galina Tremper. Weakly supervised learning of presupposition relations between verbs. In *Proceedings of the ACL 2010, Student Research Workshop*, pages 97–102, Uppsala, Sweden, 2010.

Rob van der Sandt. Presupposition projection as anaphora resolution. *Journal of Semantics*, 9:333–377, 1992.

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