Automatic Acquisition of the *Argument-Predicate* Relations from a Frame-Annotated Corpus

Ekaterina Ovchinnikova
University of Osnabrück
eovchinn@uos.de

Theodore Alexandrov
University of Bremen
theodore@math.uni-bremen.de

Tonio Wandmacher
University of Osnabrück
twandmac@uos.de

Abstract

This paper presents an approach to automatic acquisition of the argument-predicate relations from a semantically annotated corpus. We use SALSA, a German newspaper corpus manually annotated with role-semantic information based on frame semantics. Since the relatively small size of SALSA does not allow to estimate the semantic relatedness in the extracted argument-predicate pairs, we use a larger corpus for ranking. Two experiments have been performed in order to evaluate the proposed approach. In the first experiment we compare automatically extracted argument-predicate relations with the gold standard formed from associations provided by human subjects. In the second experiment we calculate correlation between automatic relatedness measure and human ranking of the extracted relations.

1 Introduction

There are many debates in lexical semantics about what kind of world knowledge actually belongs to the meaning of a lexeme. Nowadays, it is widely accepted that predicates impose selectional restrictions on their arguments. For example, since we know that the predicate *to be hungry* mainly takes expressions describing animate beings as arguments, we can correctly resolve the anaphora in the following sentence: *We gave the bananas to the monkeys because they were hungry*. There exists also multiple linguistic evidence showing that the semantics of arguments can help to predict implicit predicates. For example, the sentence *John finished the cigarette* usually means *John finished smoking the cigarette* because the meaning of the noun *cigarette* is strongly associated with the smoking activity.

It has been claimed that information about predicates associated with nouns can be helpful for a variety of tasks in natural language processing (NLP), see for example (Pustejovsky et al., 1993; Voorhees, 1994). However, at present there exists no corresponding lexical semantic resource. Several approaches have been presented that aim at creating a knowledge base containing noun-verb relations. There are two main research paradigms for developing such knowledge bases. The first paradigm assumes manual development of the resource (Pustejovsky et al., 2006), while the second one relies on automatic acquisition methods, see for example (Cimiano and Wenderoth, 2007). In this paper we propose a procedure for automatic acquisition of argument-predicate relations from a semantically annotated corpus. In line with (Lapata and Lascarides, 2003) our approach is based on the assumption that predicates are omitted in a discourse when they are highly predictable from the semantics of their arguments. We exploit SALSA (Burchardt et al., 2006), a German newspaper corpus manually annotated with FrameNet frames based on frame semantics. Using a manually annotated corpus for relation extraction has one particular advantage compared to extraction from plain text: the type of an argument-predicate relation is already annotated; there is no need to determine it by automatic means which are usually error-prone. However, the relatively small size of SALSA does not allow to make relevant predictions about the degree of semantic relatedness in the extracted argument-predicate pairs, see section 4. We therefore employ a considerably larger unannotated corpus for weighting. The results are evaluated quantitatively against human judgments obtained experimentally. The proposed evaluation procedure is similar to that presented in (Cimiano and Wenderoth, 2007). First, we create a gold standard for 30 words from the argument list and evaluate our approach with respect to this

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1388–1397,
Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

gold standard. Second, we provide results from an evaluation in which test subjects are asked to rate automatically extracted relations using a four-point scale.

The paper is structured as follows: Section 2 describes some linguistic phenomena requiring inferences of an implicit predicate from the semantics of an explicitly given argument. In section 3 we give a short overview of the related work. Sections 4 discusses the SALSA corpus. Section 5 introduces our approach. Finally, section 6 describes an experimental evaluation of the presented approach and section 7 concludes the paper.

2 Implicit Predicates

In this section we discuss some linguistic phenomena requiring inferences of an implicit predicate from the semantics of an explicitly given argument for their resolution. One of the most studied phenomena that Pustejovsky (1991) has called logical metonymy is illustrated by the examples (1a) and (1b) below. In the case of logical metonymy an implicit predicate is inferable from particular verb-noun and adjective-noun pairs in a systematic way. The verb *anfangen* ’to start’ and the adjective *kompliziert* ’complicated’ in the mentioned examples semantically select for an event, while the nouns (*Buch* ’book’ and *Frage* ’question’ respectively) have a different semantic type. However, the set of the most probable implicit predicates is predictable from the semantics of the nouns. Thus, (1a) plausibly means *Als ich angefangen habe, dieses Buch zu lesen/schreiben...* ’When I have started to read/write this book...’ and (2a) plausibly means *eine Frage die kompliziert zu beantworten ist* ’a question which is complicated to answer’.

Example 1

(a) *Als ich mit diesem Buch angefangen habe...*
’When I have started this book...’

(b) *eine komplizierte Frage*
’a complicated question’

(c) *Studentenfutter*
’student food’

(d) *Nachrichtenagentur Xinhua über Beziehungen beider Seiten der Taiwan-Strasse*
’News agency Xinhua about relations of both sides of the Taiwan Strait’

(e) *Hans ist beredt*
’Hans is eloquent’

As we can see from Example 1, besides logical metonymy there are other linguistic phenomena requiring knowledge about predicates associated with an argument for their resolution. Example (1c) contains a noun compound which can be interpreted on basis of the meaning of the noun *Futter* ’food’. In general, noun compounds can be interpreted in many different ways depending on the semantics of the constituencies: *morning coffee* is a coffee which is drunk in the morning, *brick house* is a house which is made of bricks etc. In case of (1c) the relation via the predicate *essen* ’to eat’ taking *Studenten* ’students’ as a subject and *Futter* ’food’ as an object seems to be the most plausible one.

The phrase (1d) is a title of a newspaper article. As in the previous examples, a predicate is left out in (1d). The meaning of the preposition *über* ’about’ can help to narrow down the set of possible predicates, but still allows an inadequately large range of interpretations. However, the semantics of the noun *Nachrichtenagentur* ’news agency’ supports such interpretations as *berichten* ’to report’ or *informieren* ’to inform’.

Most of the literature discusses predicates inferable from nouns. However, other parts of speech can support similar inferences. In example (1e) a predicate is predictable on the basis of the meaning of the adjective *beredt* ’eloquent’. The sentence (1e) most plausibly means that Hans speaks eloquently.

Example 1 shows that knowledge about predicates associated with explicitly given arguments can help to deal with several linguistic phenomena. The cases when a predictable predicate is left out are not rare in natural language. For example, for logical metonymy a corpus study has shown that the constructions like *begin V NP* occur rarely if the verb *V* corresponds to a highly plausible interpretation of *begin NP* (Briscoe et al., 1990).

3 Related Work

The most influential account of logical metonymy is provided by Pustejovsky’s theory of the Generative Lexicon, GL (Pustejovsky, 1991). According to Pustejovsky the meaning of a noun includes a *qualia structure* representing “the essential attributes of an object as defined by the lexical item”. Thus, the lexical meaning of the noun *book* includes *read* and *write* as qualia roles. In the framework of GL, Pustejovsky et al. (2006)

are manually developing the Brandeis Semantic Ontology which is a large generative lexicon ontology and dictionary. There also exist several approaches to automatic acquisition of qualia structures from text corpora which aim at supporting the time-consuming manual work. For example, Pustejovsky et al. (1993) use generalized syntactic patterns for extracting qualia structures from a partially parsed corpus. Cimiano and Wenderoth (2007) suggest a pattern-based method for automatic extraction of qualia structures from the Web. The results of the human judgment experiment reported in (Cimiano and Wenderoth, 2007) suggest that the automatic acquisition of qualia structures is a difficult task. Human test subjects have shown a very low agreement (11,8% average agreement) in providing qualia structures for given nouns.

Another line of research on inferring implicit predicates concerns using information about collocations derived from corpora. For example, Lapata and Lascarides (2003) resolve logical metonymy on the basis of the distribution of paraphrases like *finish the cigarette – finish smoking the cigarette* and *easy problem – problem which is easy to solve* in a corpus. This approach shows promising results, but it is limited to logical metonymy. Similarly, Nastase et al. (2006) use grammatical collocations for defining semantic relations between constituents in noun compounds.

In our study we aim at extracting intuitively plausible argument-predicate relations from a semantically annotated corpus. Using an annotated corpus we avoid problems of defining types of these relations by automatic means which are usually error-prone. We represent argument-predicate relations in terms of FrameNet frames which allow for a fine-grained and grounded representation supporting paraphrasing, see next sections. Our approach is not restricted to nouns. We also concern relations where argument positions are filled by adjectives, adverbs or even verbs.

4 The SALSA Corpus

For relation extraction we have chosen the SALSA corpus (Burchardt et al., 2006) developed at Saarland University. SALSA is a German corpus manually annotated with role-semantic information, based on the syntactically annotated TIGER newspaper corpus (Brants et al., 2002). The 2006 SALSA release which we have used contains about 20 000 annotated predicate instances. The corpus is annotated with the set of FrameNet frames.

The FrameNet, FN (Ruppenhofer et al., 2006), lexical resource is based on frame semantics (Fillmore, 1976), see *http://framenet.icsi.berkeley.edu*. The lexical meaning of predicates in FN is expressed in terms of frames (approx. 800 frames) which are supposed to describe prototypical situations spoken about in natural language. Every frame contains a set of roles (or frame elements, FEs) corresponding to the participants of the described situation. Predicates with similar semantics are assigned to the same frame, e.g. *to give* and *to hand over* refer to the GIVING frame. Consider a FN annotation for the sentence (2a) below. In this annotation DONOR, RECIPIENT and THEME are roles in the frame GIVING and *John*, *Mary* and *a book* are fillers of these roles. The FN annotation generalizes across near meaning-preserving transformations, see (2b).

Example 2

(a) $[John]_{\text{DONOR}}$ $[gave]_{\text{GIVING}}$ $[Mary]_{\text{RECIPIENT}}$ $[a\ book]_{\text{THEME}}$.

(b) $[John]_{\text{DONOR}}$ $[gave]_{\text{GIVING}}$ $[a\ book]_{\text{THEME}}$ $[to\ Mary]_{\text{RECIPIENT}}$.

In FN information about syntactic realization patterns of frame elements as well as information about frequency of occurrences of these patterns in corpora is provided. For example, the role DONOR in the frame GIVING is most frequently filled by a noun phrase in the subject position or by a prepositional phrase with the preposition *by* as the head in the complement position.

The FN project originally aimed at developing a frame-semantic lexicon for English. Later on FN frames turned out to be to a large extent language independent (Burchardt et al., 2006). In most of the cases German predicates could be successfully described by the FN frames. However, some of the frames required adaptation to the German data, e.g. new FEs were introduced. Since FN does not cover all possible word senses, new frames needed to be added for some of the predicates.

We have chosen the SALSA corpus for our experiments because to our knowledge it is the only freely available corpus which contains both syntactic and role-semantic annotation. However, we are aware that SALSA (approx. 700 000 tokens) is too small to compute a reliable co-occurrence model for measuring plausibility of the extracted argument-predicate relations, though it is relatively large for a manually annotated corpus. As it was shown in (Bullinaria and Levy, 2007), co-occurrence-based approaches need very large training corpora in order to reliably compute semantic relatedness. The SALSA corpus, comprising less than 1 million tokens, is too small for this purpose. Moreover, a considerable number of predicates in SALSA appeared to be unannotated. Some of the high frequency pairs, as for example *Bombe, explodieren* 'bomb, to explode', occur in SALSA only once, just as occasional pairs like *Deutsche, entdecken* 'German, to discover'. We have tried to overcome the size problems by using a larger unannotated corpus for recomputing the rating of our resulting relations, see next section.

5 Automatic Acquisition of the *Argument-Predicate* Relations

In line with (Lapata and Lascarides, 2003), our approach to extraction of argument-predicate (AP) relations is based on two assumptions:

A1: If predicates are highly predictable from the semantics of their arguments then they can be omitted in a discourse;
A2: If a predicate frequently takes a word as an argument then it is highly predictable from the semantics of this word.

In the proposed experimental setting argument-predicate relations are defined in terms of the FrameNet frames. Thus, we aim at extracting from SALSA tuples of the form ⟨*Argument*, ROLE, FRAME, *Predicate*⟩ such that the *Argument* plausibly fills the ROLE in the FRAME evoked by the *Predicate*. As already mentioned in section 3, our approach is not restricted to nouns. We also treat arguments expressed by other content parts of speech. The proposed relation extraction procedure consists in

- finding for every content word which occurs in the corpus a set of predicates taking this word as an argument with a high probability;
- defining a relation between the word and every predicate from this set by finding which roles the noun fills in frames evoked by the predicate;
- estimating the degree of the semantic relatedness in the extracted argument-predicate pairs.

For example, analyzing the following sentence

[*Fünf Oppositionelle*]$_{\text{SUSPECT}}$ *sind in Ebebiyin* [*von der Polizei*]$_{\text{AUTHORITIES}}$ [*festgenommen*]$_{\text{ARREST}}$ *worden*.

'Five members of the opposition have been arrested by the police in Ebebiyin.'

we aim at extracting the following tuples:

Argument	**Role**	**Frame**	**Predicate**
Oppositionell	SUSPECT	ARREST	*festnehmen*
Polizei	AUTHORITIES	ARREST	*festnehmen*

Relation Extraction

In SALSA, every sentence is annotated with a set of frames in such a way that for every frame its FEs refer to some syntactic constituents in the sentence. In order to extract argument-predicate relations from SALSA we need 1) to find a content head for every constituent corresponding to a FE; 2) to resolve possibly existing anaphora. Since SALSA is syntactically annotated, the first task proved to be relatively easy.[1] On the contrary, anaphora resolution is well-known to be one of most challenging NLP tasks. In our study, we do not focus on it, and we treat only pronominal anaphora using the following straightforward resolution algorithm: given a pronoun the first noun which agrees in number and gender with the pronoun is supposed to be its antecedent. In order to evaluate this resolution procedure we have inspected 100 anaphoric cases. In approximately three fourths of the cases the anaphora were resolved correctly. Therefore, we have assigned a confidence rate of 0,75 to the FE fillers resulting from a resolved anaphora. In non-anaphoric cases a confidence rate of 1 was assigned.

For every extracted tuple of the form ⟨*Argument*, ROLE, FRAME, *Predicate*⟩ we have summed up the corresponding confidence rates. Finally, we have obtained around 30 000 tuples with confidence rates ranging from 0,75 to 88. It is not surprising that most of the arguments appeared to be nouns, while most of the predicates are expressed by verbs. Since SALSA has been annotated manually, there are almost no mistakes in defining types of the semantic

[1] We have excluded from the consideration foreign-language expressions, while proper nouns were treated in the usual way. For verb phrases with auxiliary or modal verbs as heads the main verb was taken as a corresponding role filler.

relations between arguments and predicates.[2] For several pairs, the semantic relation between an argument and a predicate is ambiguous. Consider the tuples extracted for the word pair *Buch, schreiben* 'book, to write' which are given below. While the first tuple corresponds to phrases like *ein Buch schreiben* 'to write a book', the second one abstracts from the expressions like *in einem Buch schreiben* 'to write in a book'.

Argument	**Role**	**Frame**	**Predicate**
Buch	TEXT	TEXT_CREATION	*schreiben*
Buch	MEDIUM	STATEMENT	*schreiben*

Additionally, ambiguity can arise because of the annotation disagreements in SALSA. For example, the pair (*Haft*, *sitzen*) 'imprisonment', 'to sit' in Table 1 was annotated in SALSA both with the BEING_LOCATED and with the POSTURE frames.

As mentioned in section 4, a considerable number of predicates in SALSA is not annotated semantically. In order to find out how many relevant AP-relations get lost if we consider only semantically annotated predicates, we have additionally extracted AP-pairs on the basis of the syntactic annotation only. The anaphora resolution procedure as described above was again applied to the syntactic argument heads. We have obtained around 56 500 pairs with confidence rates ranging from 0,75 to 71,50.[3]

As one could expect, being a newspaper corpus SALSA appeared to be thematically unbalanced. The most frequent argument-predicate relations occurring in SALSA reflect common topics discussed in newspapers: economics (e.g. (*Prozent, steigen*), 'percent', 'to increase'), criminality (e.g. (*Haft, verurteilen*) 'imprisonment', 'to sentence'), catastrophes (e.g. (*Mensch, töten*) 'human', 'to kill') etc.

Ranking

As mentioned in section 4, the size of SALSA does not allow to make relevant predictions about the distribution of frames and role fillers. Only 2% of the relations occur in SALSA more then 3 times. In order to overcome this problem we have developed a measure of semantic relatedness between the extracted arguments and predicates which takes into account their co-occurrence in a larger and more representative corpus. For computing semantic relatedness we have used a lemmatized newspaper corpus (Süddeutsche Zeitung, SZ) of 145 million words. Given a tuple t with a confidence rate c containing an argument a and a predicate p, the relatedness measure *rm* of t was computed as follows:

$$rm(t) = lsa(a,p) + c/max(c),$$

where the $lsa(a,p)$ is based on Latent Semantic Analysis, LSA (Deerwester et al., 1990). LSA is a vector-based technique that has been shown to give reliable estimates on semantic relatedness. It makes use of distributional similarities of words in text and constructs a semantic space (or word space) in which every word of a given vocabulary is represented as a vector. Such vectors can then be compared to one another by the usual vector similarity measures (e.g. cosine). We calculated the LSA word space using the Infomap toolkit10 v. 0.8.6 (*http://infomap-nlp.sourceforge.net*). The co-occurrence matrix (window size: 5 words) comprised 80 000×3 000 terms and was reduced by SVD to 300 dimensions. For the vector comparisons the cosine measure was applied. To those words which did not occur in the analyzed SZ corpus (approx. 3500 words) a *lsa* measure of 0 was assigned. To provide a comparable contribution to rm, the confidence rates c extracted from SALSA are divided by the maximal confidence rate. The *rm* function is a linear interpolation of the *lsa* and the normalized *c* measure. As mentioned above, the *c* measure is a discriminative factor for only 2% of the relations. For the remaining 98% the normalized *c* values are small (0,003 or 0,002 or 0,001). Therefore, calculating the *rm* measure we mainly rely on *lsa*, while normalized *c* actually plays a role only for the relations frequently occurring in SALSA. Table 1 contains the 5 most semantically related predicates for an example argument.

6 Evaluation

Since the extracted argument-predicate relations are intended to be used for inferring intuitively obvious predicates,we evaluate to which extend they correspond to human intuition.

[2] Mistakes can arise only because of the annotation errors and errors in the anaphora resolution procedure.

[3] The comparison of the results obtained by the extraction procedure based on the semantic annotation with the results of the procedure based on the syntactic annotation only is provided in the next section.

Table 1: Examples of the extracted argument-predicate relations

Argument	**Role**	**Frame**	**Predicate**	***rm***
Haft 'imprisonment'	FINDING	VERDICT	*verurteilen* 'to sentence'	0,939
	LOCATION	BEING_LOCATED	*sitzen* 'to sit'	0,237
	LOCATION	POSTURE	*sitzen* 'to sit'	0,226
	MESSAGE	REQUEST	*fordern* 'to demand'	0,153
	BAD_OUTCOME	RUN_RISK-FNSALSA	*drohen* 'to threaten'	0,144

Gold Standard

Similar to (Cimiano and Wenderoth, 2007) we provide a gold standard for 30 test arguments occurring in the SALSA corpus. The test arguments were selected randomly from the set of those arguments that have more than one predicate associated with them such that a value of argument-predicate relatedness exceeds the average one. These words were nearly uniformly distributed among 20 participants of the experiment, who were all non-linguists. We also ensured that each word was treated by three different subjects. For every word we asked our subjects to write between 5 and 10 short phrases that contain a predicate taking the given word as an argument, e.g. *book – to read a book.* The participants were asked to provide phrases instead of single predicates, because we wanted to control the syntactic and semantic position of the arguments. The participants received an instruction informally describing the notion of predicate and what kind of phrases they are supposed to come up with. Besides the task description they were shown examples containing appropriate and inappropriate phrases. Some of the examples are given below.

Example 3

(a) *Aktie* 'stock' : *Kauf der Aktien* 'buying of stocks', *Aktien kaufen* 'to buy stocks', *Aktien an der Börse* 'stocks on the bourse' (is inappropriate because the word "bourse" describes a place and not an event)

(b) *beredt* 'eloquent': *beredt sprechen* 'to speak eloquently', *ein beredter Sprecher* 'an eloquent speaker' (is inappropriate because the word "speaker" describes a person and not an event)

The test was conducted via e-mail. In order to compare the human associations with the extracted AP-relations, we have manually annotated the obtained phrases with SALSA frames. The agreement for the described task for every cue word was calculated as the averaged pairwise agreement between the AP-relations delivered by the three subjects, S_1, S_2 and S_3, as follows:

$$Agr = \frac{\frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} + \frac{|S_2 \cap S_3|}{|S_2 \cup S_3|} + \frac{|S_2 \cap S_3|}{|S_2 \cup S_3|}}{3}.$$

Agreement results for every cue word are reported in table 2. Second column of the table contains gold standard predicates which were provided by all 3 participants treating the same word.[4] Averaging over all words, we got a mean agreement of 13%. Though this value seems to be low, it is consistent with a mean agreement of 11,8% for a similar task reported in (Cimiano and Wenderoth, 2007), see section 3. Cimiano and Wenderoth (2007) show that the lowest agreement is yielded for more abstract words, while the agreement for very concrete words is reasonable. We could not make a similar observation, see table 2.

Comparison with the Gold Standard

In the first experiment we checked whether predicates which people associate with the test arguments can be automatically extracted by our procedure. For this aim we compared the gold standard with all automatically extracted argument-predicate relations[5] containing some of the 30 cue words as follows. These relations were ranked according to the relatedness measure described in previous section. In line with (Cimiano and Wenderoth, 2007) we exploited an approach common in information retrieval for estimating the quality of correspondence of a ranked output to a gold standard, see (Baeza-Yates and Ribeiro-Neto, 1999).

Given some n automatically extracted relations with the highest ranking we calculated a precision-recall curve expressing precision and recall of our procedure compared to the gold standard. The precision characterizes the procedure exactness, i.e. how many redundant relations are retrieved. The

[4] The overall gold standard consists of 33 tuples.

[5] In order to evaluate the procedure extracting AP-relations on the basis of the semantic annotation we compared automatically extracted tuples to the gold standard tuples. For the procedure using the syntactic annotation only the AP-pairs were considered without regarding frames and FEs.

recall measures the completeness, i.e. how many relations of the gold standard are extracted automatically. For each point of the curve (which is a pair (p, r) of values of precision p and recall r) we calculated the F-measure as $F = 2pr/(p + r)$ which is the harmonic mean between recall and precision. The precision-recall curve is a set of precision values for the prespecified recall levels varying from 0 to 1 with a step 0,1. Then, to produce only one value evaluating the quality of the ranked output compared to the gold standard, for each precision-recall curve we calculated F_{max}, the maximal value of the F-measure achieved for the points of this curve. F_{max} expresses the best trade-off between precision and recall for the given ranked output. Finally, among all possible n (numbers of the considered relations with the highest ranking) we selected that one which provides the maximal F_{max} value.

The resulting maximal F_{max} values are 0,47 for the procedure extracting AP-relations on the basis of the semantic annotation and 0,41 for the procedure using the syntactic annotation only. We compared these results with the baseline results of maximal F_{max} values produced for the output with random ranking. The calculation of the baseline was repeated 100 times, each time a new random ranking was generated. The lowest baseline results are 0,08/0,06 (semantic/syntactic annotation), the highest are 0,18/0,14 and the medians are 0,1/0,07. One can see that the results produced using the relatedness measure (0,47/0,41) greatly exceed the baseline. Based on this comparison we conclude that the ranking done using the relatedness measure brings a significant advantage. The values of precision and recall for the reported maximal F_{max} values are 0,5/0,33 (semantic/syntactic annotation) and 0,45/0,54 respectively. This results show that half of the AP-relations from the gold standard appeared to be in the list of the top-ranked tuples extracted by the "semantic" procedure, while the size of this list ($n = 28$) was almost equal to the size of the gold standard (33). The differences in performance between the "semantic" and "syntactic" procedures could be explained by the fact that the "syntactic" procedure finds in the corpus more related predicates for every argument than the "semantic" one. Nevertheless, the "semantic" procedure shows better performance.

Next we investigated the results for each argument used in the gold standard separately in the same way as described above. For each argument the F_{max} measure has been computed. Because of the low agreement between the subjects questioned for the gold standard (see above), in these calculations we considered all predicates reported by our subjects. The calculated F_{max} values are reported in table 2 which shows a correlation between F_{max} values calculated for the "semantic" and "syntactic" procedures. However, there is no correlation with human agreement. This issue needs a further investigation, see section 7.

Human Judgments of the Relatedness

Following (Cimiano and Wenderoth, 2007), in order to check whether the calculated relatedness is reasonable according to human intuition, we have performed another experiment. For each of the 30 words selected for the gold standard we selected the 5 top ranked predicates. Since for some of the cue arguments only 3 predicates were found in the corpus, the final test set contains only 138 argument-predicate tuples. From these tuples we generated short grammatically correct phrases structurally similar to those in example 3. These phrases were uniformly distributed among 10 subjects so that every phrase was evaluated by one subject. The participants were asked to rate the phrases with respect to their naturalness using a scale from 0 to 3, whereby 0 means 'unnatural', 1 'possible', 2 'natural' and 3 'totally natural and self-evident'.

Further on we investigated the relationship between the human estimates and the relatedness values obtained automatically. For this aim we used the Spearman rank correlation coefficient. Because of four-points scale used, the human rankings are equal for many tuples which lead to the so-called effect of ties. For this reason we computed the correlation coefficient with a correction for ties. The coefficient value is 0,30 and this correlation is statistically significant with p-value 0,0006. Based on these results we conclude that our relatedness measure is correlated with human judgments. Taking into account the subjective character of human ranking in terms of naturalness, the achieved correlation values can be considered as high.

Table 2: Evaluation results for 30 gold standard cue words.

Cue word	**Shared predicates**	***Agr***	**Sem.** F_{max}	**Syn.** F_{max}
Name ’name’	*haben* ’to have’	14%	0,2	0,48
Urlaub ’vacation’	*fahren* ’to go’	8%	0,13	0,16
Sprache ’language’	*sprechen* ’to speak’, *lernen* ’to learn’	14%	0,4	0,3
Strafe ’fine’	*verurteilen* ’to sentence’	11%	0,21	0,3
Stuhl ’chair’	*sitzen* ’to sit’	14%	0,1	0,2
Bombe ’bomb’	*hochgehen* ’to blow up’	14%	0,11	0,22
Blatt ’gazette’, ’page’, ’leaf’	–	2%	0	0
Flughafen ’airport’	*ankommen* ’to arrive’, *fahren* ’to go’	21%	0,17	0,17
Gesetz ’low’	–	8%	0,17	0,38
Polizei ’police’	*rufen* ’to call’	11%	0,22	0,23
Kompromiss ’compromise’	*schliessen* ’to make’	15%	0,07	0,29
Fluggesellschaft ’airline’	–	3%	0,11	0,38
Antrag ’proposal’, ’application’	*stellen* ’to introduce’, *ablehnen* ’to decline’	24%	0,43	0,42
Zeitung ’newspaper’	*lesen* ’to read’	13%	0,17	0,09
Brief ’letter’	*verschicken* ’to send’, *schreiben* ’to write’	19%	0,23	0,12
Flüchtling ’refugee’	*aufnehmen* ’to accept’	13%	0	0,07
Buch ’book’	*schreiben* ’to write’, *lesen* ’to read’	15%	0,44	0,39
Zähler ’counter’	*ablesen* ’ to read’	11%	0	0
Anzahl ’number’	–	3%	0,23	0,19
Prozent ’percent’	–	3%	0,48	0,21
Ziel ’goal’	*verfehlen* ’to miss’, *erreichen* ’to reach’	20%	0,3	0,48
Schule ’school’	*schwänzen* ’to miss’, *gehen* ’to go’	22%	0,13	0,23
Amt ’position’, ’department’	*bekleiden, innehaben* ’to hold’, *gehen* ’to go’	20%	0	0,17
Frage ’question’	*beantworten* ’to answer’, *stellen* ’to ask’	20%	0,15	0,37
Mensch ’human’	*sein* ’to be’	16%	0,09	0,03
Zeuge ’witness’	*aussagen* ’to testify’, *sein* ’to be’	22%	0,13	0,19
Thema ’theme’	–	7%	0,14	0,26
Preisträger ’prize winner’	–	5%	0,08	0,08
Initiative ’initiative’	*ergreifen* ’to take’	17%	0,1	0,13
Wohnung ’flat’	–	7%	0,09	0,17

7 Conclusion and Discussion

In this paper we presented an approach to automatic extraction of argument-predicate relations from a frame-annotated corpus.[6] In our approach we aimed to combine the advantages offered by annotated and unannotated lexical resources. Besides extracting AP-pairs the proposed method allows us to define types of semantic relations in terms of FrameNet frames. The proposed procedure is not restricted to arguments expressed by nouns and treats also other content parts of speech.

The main goal of this paper was to show that though manually annotated corpora usually have a relatively small size, they can be successfully exploited for the relation extraction. An obvious limitation of the presented approach is that it is bounded to manual annotations which are hard to obtain. However, since semantic annotations are useful for many different goals in linguistics and NLP, the number of reliable annotated corpora constantly grows.[7] Moreover, recently several tools have been developed which perform role annotation automatically, for example see (Erk and Pado, 2006). Therefore we believe that approaches using semantic annotation are valid and promising. In the future we plan to experiment with large role-annotated corpora for English such as PropBank (approx. 300 000 words, (Palmer et al., 2005)) and the FrameNet-annotated corpus provided by the FN project (more than 135 000 annotated sentences, (Ruppenhofer et al., 2006)). Since these corpora do not contain syntactic annotation, for extracting argument-predicate relations we will need to parse annotated sentences.

There are several ways to improve the proposed procedure. First, an implementation of a more advanced anaphora resolution algorithm treating pronominal as well as nominal anaphora should significantly raise the precision/recall characteristics. Second, splitting German compounds occurring in the corpus should provide additional evidence. We have treated such words as *Kunde* ’client’ and *Privatkunde* ’private client’ as different lexemes, while they are strongly related se-

[6] The complete list of the extracted AP-relations as well as the results of the experiment will be available online at *http://www.ikw.uni-osnabrueck.de/∼eovchinn/APrels/*.

[7] At present FrameNet annotated corpora are available for English, German and Spanish, see *http://framenet.icsi.berkeley.edu*.

mantically and information about predicates co-occurring with the second word could probably be used for describing the semantics of the first one. Concerning relatedness measure, additional corpus-based measures such as Web-based measures (Cimiano and Wenderoth, 2007) or measures based on syntactic relations (Pustejovsky et al., 1993) could appear to be useful for improving the ranking of the extracted relations.

The presented procedure was evaluated quantitatively against human judgments obtained experimentally. The participants of the experiment were asked to provide short phrases containing given cue words and predicates associated with these words as well as to rate phrases generated from the automatically extracted AP-relations. Concerning the first experiment, the low human agreement has shown that the proposed association task appeared to be difficult for the subjects. Nevertheless, the described learning procedure proved to extract intuitively reasonable relations.

The evaluation strategy presented in this paper on relies on the underlying assumptions (**A1** and **A2** in section 5) and is compatible with the other approaches to relation extraction, cf. (Cimiano and Wenderoth, 2007). However, it is plausible that human responses in the context of providing associated predicates for target words will differ from the responses in the experimental settings where subjects are asked to infer implicit predicates, e.g. to extend phrases containing implicit predicates. In the future we plan to implement a procedure making use of the extracted AP-relations which would automatically extend phrases containing implicit predicates. Then we intend to compare output results of the procedure with the human responses. Additionally, a study of a possible correspondence between human agreement on associated predicates and a semantic type of an argument (e.g. concrete/abstract, natural kind/artifact) should be performed on more test arguments.

Potential Applications

As already mentioned in the literature, see for example (Lapata and Lascarides, 2003), knowledge about implicit predicates could be potentially useful for a variety of NLP tasks such as language generation, information extraction, question answering or machine translation. Many applications of semantic relations in NLP are connected to paraphrasing or query expansion, see for example (Voorhees, 1994). Suppose that a search engine or a question answering system receives the query *schnelle Bombe* 'quick bomb'. Probably, in this case the user is interested in finding information about bombs that explode quickly rather then about bombs in general. Knowledge about predicates associated with the noun *Bombe* 'bomb' could be used for predicting a set of probable implicit predicates. For generation of the semantically and syntactically correct paraphrases it is sometimes not enough to guess the most probable argument-predicate pairs. Information about types of an argument-predicate relation could be helpful, i.e. which semantic and syntactic position does the argument fill in the argument structure of the predicate. For example, compare *eine Bombe explodiert schnell* 'a bomb explodes quickly' for *schnelle Bombe* with *ein Buch schnell lesen/schreiben* 'to read/write a book quickly' for *schnelles Buch* 'quick book'. In the first case the argument *Bombe* fills the subject position, while in the second case *Buch* fills the object position. Since FrameNet contains information about syntactic realization patterns for frame elements, representation of argument-predicate relations in terms of frames directly supports generation of semantically and syntactically correct paraphrases.

The described procedure could also support manual development of a lexical resource, providing evidence from corpora as well as the distributional information.

References

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison Wesley, Harlow, 1. aufl. edition.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*.

Ted Briscoe, Ann Copestake, and Bran Boguraev. 1990. Enjoy the paper: Lexical semantics via lexicology. In *Proceedings of the 13th International Conference on Computational Linguistics*, pages 42–47.

John Bullinaria and Joseph Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal.

2006. The SALSA corpus: A German corpus resource for lexical semantics. In *Proceedings of LREC 2006*, pages 969–974.

Philipp Cimiano and Johanna Wenderoth. 2007. Automatic Acquisition of Ranked Qualia Structures from the Web. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 888–895.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *American Society of Information Science*, 41(6):391–407.

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

Charles J. Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32.

Mirella Lapata and Alex Lascarides. 2003. A Probabilistic Account of Logical Metonymy. *Computational Linguistics*, 29(2):261–316.

Vivi Nastase, Jelber Sayyad-Shirabad, Marina Sokolova, and Stan Szpakowicz. 2006. Learning noun-modifier semantic relations with corpus-based and wordnet-based features. In *Proceedings of the AAAI 2006*.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The Proposition Bank: A Corpus Annotated with Semantic Roles. *Computational Linguistics*, 31(1):71–106.

James Pustejovsky, Peter Anick, and Sabine Bergler. 1993. Lexical semantic techniques for corpus analysis. *Computational Linguistics*, 19(2):331–358.

James Pustejovsky, Catherine Havasi, Roser Saur, Patrick Hanks, Anna Rumshisky, Jessica Littman, Jos Castao, and Marc Verhagen. 2006. Towards a generative lexical resource: The Brandeis Semantic Ontology. In *Proceedings of the Fifth Language Resource and Evaluation Conference*.

James Pustejovsky. 1991. The Generative Lexicon. *Computational Linguistics*, 17(4):409–441.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. FrameNet II: Extended Theory and Practice. *International Computer Science Institute*.

Ellen M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69.