Automatically Identifying Implicit Arguments to Improve Argument Linking and Coherence Modeling

Michael Roth and **Anette Frank**
Department of Computational Linguistics
Heidelberg University, Germany
{mroth,frank}@cl.uni-heidelberg.de

Abstract

Implicit arguments are a discourse-level phenomenon that has not been extensively studied in semantic processing. One reason for this lies in the scarce amount of annotated data sets available. We argue that more data of this kind would be helpful to improve existing approaches to linking implicit arguments in discourse and to enable more in-depth studies of the phenomenon itself. In this paper, we present a range of studies that empirically validate this claim. Our contributions are threefold: we present a heuristic approach to automatically identify implicit arguments and their antecedents by exploiting comparable texts; we show how the induced data can be used as training data for improving existing argument linking models; finally, we present a novel approach to modeling local coherence that extends previous approaches by taking into account non-explicit entity references.

1 Introduction

Semantic role labeling systems traditionally process text in a sentence-by-sentence fashion, constructing local structures of semantic meaning (Palmer et al., 2010). Information relevant to these structures, however, can be non-local in natural language texts (Palmer et al., 1986; Fillmore, 1986, inter alia). In this paper, we view instances of this phenomenon, also referred to as *implicit arguments*, as elements of discourse. In a coherent discourse, each utterance focuses on a *salient* set of entities, also called "foci" (Sidner, 1979) or "centers" (Joshi and Kuhn, 1979). According to the theory of Centering (Grosz et al., 1995), the salience of an entity in a discourse is reflected by linguistic factors such as choice of referring expression and syntactic form. Both extremes of salience, i.e., contexts of referential continuity (Brown, 1983) and irrelevance, can also be reflected by the non-realization of an entity. Altough specific instances of non-realization, so-called zero anaphora, have been well-studied in discourse analysis (Sag and Hankamer, 1984; Tanenhaus and Carlson, 1990, inter alia), this phenomenon has widely been ignored in computational approaches to entity-based coherence modeling. It could, however, provide an explanation for local coherence in cases that are not covered by current models of Centering (cf. Louis and Nenkova (2010)). In this work, we propose a new model to predict whether realizing an argument contributes to local coherence in a given position in discourse. Example (1) shows a text fragment, in which argument realization is necessary in the first sentence but redundant in the second.

(1) El Salvador is now the only Latin American country which still has troops in [Iraq]. Nicaragua, Honduras and the Dominican Republic have withdrawn their troops [∅].

From a semantic processing perspective, a human reader can easily infer that "Iraq", the marked entity in the first sentence of Example (1), is also an implicit argument of the predicate "withdraw" in the second sentence. This inference step is, however, difficult to model computationally as it involves an interplay of two challenging sub-tasks: first, a semantic processor has to determine that an argument is not realized (but inferrable); and second, a suit-

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 306–316, Atlanta, Georgia, June 13-14, 2013.

able antecedent has to be found within the discourse context. For the remainder of this paper, we refer to these steps as *identifying* and *linking* implicit arguments to discourse antecedents.

As indicated by Example (1), implicit arguments are an important aspect in semantic processing, yet they are not captured in traditional semantic role labeling systems. The main reasons for this are the scarcity of annotated data, and the inherent difficulty of inferring discourse antecedents automatically.

In this paper, we propose to induce implicit arguments and discourse antecedents by exploiting complementary (explicit) information obtained from monolingual comparable texts (Section 3). We apply the empirically acquired data in argument linking (Section 4) and coherence modeling (Section 5). We conclude with a discussion on the advantages of our data set and outline directions for future work (Section 6).

2 Related work

The most prominent approach to entity-based coherence modeling nowadays is the entity grid model by Barzilay and Lapata (2005). It has originally been proposed for automatic sentence ordering but has also been applied in coherence evaluation and readability assessment (Barzilay and Lapata, 2008; Pitler and Nenkova, 2008), and story generation (McIntyre and Lapata, 2009). Based on the original model, a few extensions have been proposed: for example, Filippova and Strube (2007) and Elsner and Charniak (2011b) suggested additional features to characterize semantic relatedness between entities and features specific to single entities, respectively. Other entity-based approaches to coherence modeling include the pronoun model by Charniak and Elsner (2009) and the discourse-new model by Elsner and Charniak (2008). All of these approaches are, however, based on explicitly realized entity mentions only, ignoring references that are inferrable.

The role of implicit arguments has been studied early on in the context of semantic processing (Fillmore, 1986; Palmer et al., 1986). Yet, the phenomenon has mostly been ignored in semantic role labeling. First data sets, focusing on implicit arguments, have only recently become available: Ruppenhofer et al. (2010) organized a SemEval shared task on "linking events and participants in discourse", Gerber and Chai (2012) made available implicit argument annotations for the NomBank corpus (Meyers et al., 2008) and Moor et al. (2013) provide annotations for parts of the OntoNotes corpus (Weischedel et al., 2011). However, these resources are very limited: The annotations by Moor et al. and Gerber and Chai are restricted to 5 and 10 predicate types, respectively. The training set of the SemEval task contains only 245 resolved implicit arguments in total. As pointed out by Silberer and Frank (2012), additional training data can be heuristically created by treating anaphoric mentions as implicit arguments. Their experimental results showed that artificial training data can indeed improve results, but only when obtained from corpora with manual semantic role annotations (on the sentence level) and gold coreference chains.

3 Identifying and linking implicit arguments

The aim of this work is to automatically construct a data set of implicit arguments and their discourse antecedents. We propose an induction approach that exploits complementary information obtained from pairs of comparable texts. As a basis for this approach, we rely on several preparatory steps proposed in the literature that first identify information two documents have in common (cf. Figure 1). In particular, we align corresponding predicate-argument structures (PAS) using graph-based clustering (Roth and Frank, 2012b). We then determine co-referring entities across the texts using coreference resolution techniques on concatenated document pairs (Lee et al., 2012). These preprocessing steps are described in more detail in Section 3.1.

Given the preprocessed comparable texts and aligned PAS, we propose to heuristically identify implicit arguments and link them to their antecedents via the cross-document coreference chains. We describe the details of this approach in Section 3.2.

3.1 Data preparation

The starting point for our approach is the data set of automatically aligned predicate pairs that has been released by Roth and Frank (2012a).[1] This data

[1] cf. `http://www.cl.uni-heidelberg.de/%7Emroth/`

Sentence that comprises a PAS with an (correctly predicted) implicit argument	induced antecedent
The [$\varnothing_{A0}$] [operating$_{A3}$] **loss**, as measured by . . . widened to 189 million euros . . .	T-Online['s]
It was handed over to Mozambican control . . . 33 years after [$\varnothing_{A0}$] **independence**.	Mozambique['s]
. . . [local officials $_{A0}$] failed to immediately **report** [the accident $_{A1}$] [$\varnothing_{A2}$] . . .	[to] the government

Table 1: Three positive examples of automatically induced implicit argument and antecedent pairs.

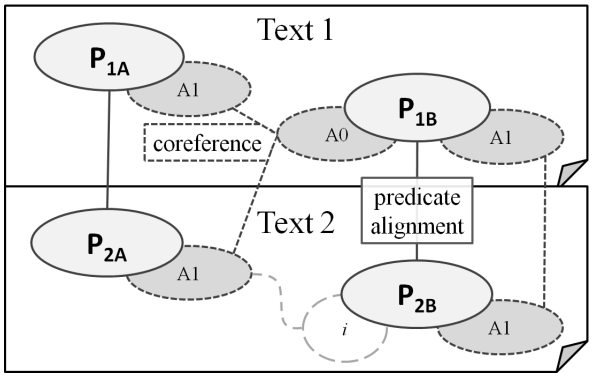

Figure 1: Illustration of the induction approach: texts consist of PAS (represented by overlapping circles); we exploit alignments between corresponding predicates across texts (marked by solid lines) and co-referring entities (marked by dotted lines) to infer implicit arguments (marked by 'i') and link antecedents (curly dashed line)

set, henceforth just **R&F data**, is a collection of 283,588 predicate pairs that have been aligned "with high precision"[2] across comparable newswire articles from the Gigaword corpus (Parker et al., 2011). To use these documents for our argument induction technique, we apply a couple of pre-processing tools on each single document and perform cross-document entity coreference on pairs of documents.

Single document pre-processing. We apply several preprocessing steps to all documents in the R&F data: we use the Stanford CoreNLP package[3] for tokenization and sentence splitting. We then apply MATE tools (Bohnet, 2010; Björkelund et al., 2010), including the integrated PropBank/NomBank-style semantic parser, to reconstruct local predicate-argument structures for aligned predicates. Finally, we resolve pronouns that occur in a PAS using the coreference resolution system by Martschat et al. (2012).

[2]The used method achieved a precision of 86.2% at a recall of 29.1% on the Roth and Frank (2012a) test set.

[3]http://nlp.stanford.edu/software/

Cross-document coreference. We apply cross-document coreference resolution to induce antecedents for implicit arguments. In practice, we use the Stanford Coreference System (Lee et al., 2013) and run it on pairs of texts by simply providing a single document as input, comprising of a concatenation of the two texts. To perform this step with high precision, we only use the most precise resolution sieves: "String Match", "Relaxed String Match", "Precise Constructs", "Strict Head Match [A-C]", and "Proper Head Noun Match".

3.2 Identification and linking approach

Given a pair of aligned predicates from two comparable texts, we examine the parser output to identify the arguments in each predicate-argument structure (PAS). We compare the set of realized argument positions in both structures to determine whether one PAS contains an argument position (explicit) that has not been realized in the other PAS (implicit). For each implicit argument, we identify appropriate antecedents by considering the cross-document coreference chain of its explicit counterpart. As our goal is to link arguments within discourse, we restrict candidate antecedents to mentions that occur in the same document as the implicit argument.

We apply a number of restrictions to the resulting pairs of implicit arguments and antecedents to minimize the impact of errors from preprocessing:

- The aligned PAS should consist of a different number of arguments (to minimize the impact of argument labeling errors)
- The antecedent should not be a resolved pronoun (to avoid errors resulting from incorrect pronoun resolution)
- The antecedent should not be in the same sentence as the implicit argument (to circumvent cases, in which an implicit argument is actually explicit but has not been recognized by the parser)

3.3 Resulting data set

We apply the identification and linking approach to the full R&F data set of aligned predicates. As a result, we induce a total of 701 implicit argument and antecedent pairs, each in a separate document, involving 535 different predicates. Examples are displayed in Table 1. Note that 701 implicit arguments from 283,588 pairs of predicate-argument structures seem to represent a fairly low recall. Most predicate pairs in the high precision data set of Roth and Frank (2012a) do, however, consist of identical argument positions (84.5%). In the remaining cases, in which an implicit argument can be identified (15.5%), an antecedent in discourse cannot always be found using the high precision coreference sieves. This does not mean that implicit arguments are a rare phenomenon in general. In fact, 38.9% of all manually aligned predicate pairs in Roth and Frank (2012a) involved a different number of arguments.

We manually evaluated a subset of 90 induced implicit arguments and found 80 discourse antecedents to be correct (89%). Some incorrectly linked instances still result from preprocessing errors. In Table 2, we present a range of different error types that occurred when extracting implicit arguments without any restrictions.

4 Experiment 1: Linking implicit arguments

Our first experiment assesses the utility of automatically induced implicit arguments and antecedent pairs for the task of implicit argument linking. For evaluation, we use the data sets from the SemEval 2010 task on Linking Events and their Participants in Discourse (Ruppenhofer et al., 2010, henceforth just **SemEval**). For direct comparison with previous results and heuristic acquisition techniques (cf. Section 2), we apply the implicit argument identification and linking model by Silberer and Frank (2012, henceforth **S&F)** for training and testing.

4.1 Task summary

Both the training and test sets of the SemEval task are text corpora extracted from Sherlock Holmes novels, with manual frame semantic annotations including implicit arguments. In the actual linking task ("NI-only"), labels are provided for local arguments and participating systems have to perform the following three sub-tasks: (1) identify implicit arguments (IA), (2) predict whether each IA is resolvable and, if so, (3) find an appropriate antecedent.

The task organizers provide two versions of their data sets: one based on FrameNet annotations and one based on PropBank/NomBank annotations. We found that the latter, however, only contains a subset of the implicit argument annotations from the FrameNet-based version. As all previous results in this task have been reported on the FrameNet data set, we adopt the same setting. Note that our additional training data is automatically labeled with a PropBank/NomBank-style parser. That is, we need to map our annotations to FrameNet. The organizers of the SemEval shared task provide a manual mapping dictionary for predicates in the annotated data set. We make use of this manual mapping and additionally use SemLink 1.1[4] for mapping predicates and arguments not in the dictionary.

4.2 Model details

We make use of the system by S&F to train a new model for the NI-only task. As mentioned in the previous sub-section, this task consists of three steps: In step (1), implicit arguments are identified as unfilled FrameNet core roles that are not competing with roles that are already filled; in step (2), a SVM classifier is used to predict whether implicit arguments are resolvable based on a small amount of features – semantic type of the affected Frame Element, the relative frequency of its realization type in the SemEval training corpus, and a boolean feature that indicates whether the affected sentence is in passive voice and does not contain a (deep) subject. In step (3), we apply the same features and classifier as S&F, i.e., the BayesNet implementation from Weka (Witten and Frank, 2005), to find appropriate antecedents for (predicted) resolvable arguments. S&F report that their best results were obtained when considering all entities as candidate antecedents that are syntactic constituents from the present and the past two sentences, or entities that occurred at least five times in the previous discourse ("Chains+Win" setting). In their evaluation, the latter of these two restrictions crucially depended on gold coreference chains. As the automatic coreference chains in our

[4] http://verbs.colorado.edu/semlink/

Sentence that comprises a PAS with an (incorrectly predicted) implicit argument	induced antecedent
(1) .. [Statistics$_*$] released [Tuesday $_{TMP}$] [$\varnothing_{A0}$] **showed** the death toll dropped ...	official statistics
(2) A [French $_{LOC*}$] [$\varnothing_{A0}$] draft **resolution** ... demands full ... compliance ...	France
(3) An earthquake ... is capable of causing .. [heavy $_{EXT}$] **damage** [$\varnothing_{A2*}$]	major

Table 2: Examples of erroneous pairs of implicit arguments and antecedents. In (1), the parser did not recognize "Statistics" as an argument of **showed**; in (2), the parser mislabeled "French" as a locative modifier; both errors lead to incorrectly identified implicit arguments. In (3), the implicit argument is correct but the wrong antecedent was identified because "major" had been mislabeled in the aligned predicate-argument structure

data are rather sparse (and noisy), we only consider syntactic constituents from the present and the past two sentences as antecedents ("SentWin" setting).

Before training and testing a new model with our own data, we perform feature selection using 10-fold cross validation. We run the feature selection on a combination of the SemEval training data and our additional data set in order to find a set of features that generalizes best across the two different corpora. We found these to be features regarding "prominence", selectional preferences ("sp_supersense"), the POS tags of entity mentions, and semantic types of argument positions ("semType_dni.entity"). Note that the S&F system does not make use of any lexicalized information. Instead, semantic features are computed based on the highest abstraction level in WordNet (Fellbaum, 1998). For detailed description of all features, see Silberer and Frank (2012).

4.3 Results

For direct comparison in the full task, both with S&F's model and other previously published results, we adopt the precision, recall and F_1 measures as defined in Ruppenhofer et al. (2010). We compare our results with those previously reported on the SemEval task (see Table 3 for a summary): Chen et al. (2010) adapted SEMAFOR, the best performing system that participated in the actual task in 2010. Tonelli and Delmonte (2011) presented a revised version of their SemEval system (Tonelli and Delmonte, 2010), which outperformed SEMAFOR in terms of recall (6%) and F_1 score (8%). The best results in terms of recall and F_1 score up to date have been reported by Laparra and Rigau (2012), with 25% and 19%, respectively. Our model outperforms their state-of-the-art system in terms of precision (21%) but at a higher cost of recall (8%). Two influencing factors for their high recall are probably (1) their improved method for identifying (resolvable) implicit arguments, and (2) their addition of lexicalized and ontological features.

	P	R	F
Chen et al. (2010)[5]	**0.25**	0.01	0.02
Tonelli and Delmonte (2011)	0.13	0.06	0.08
Laparra and Rigau (2012)	0.15	**0.25**	**0.19**
Laparra and Rigau (2013)	0.14	0.18	0.16
Gorinski et al. (2013)[6]	0.14	0.12	0.13
S&F (no additional data)	0.06	0.09	0.07
S&F (best additional data)	0.09	0.11	0.10
This paper	0.21	0.08	0.12

Table 3: Results in terms of precision (P), recall (R) and F_1 score (F) for identifying and linking implicit arguments in the SemEval test set.

Comparison to the original results reported by S&F, whose system we use, shows that our additional data improves precision (from 6% to 21%) and F_1 score (from 7% to 12%). The loss in recall is marginal (-1%) given the size of the test set (259 resolvable cases in total). The result in precision is the second highest score reported on this task. Interestingly, the improvements are higher than those of the best training set used in the original study by Silberer and Frank (2012), even though their additional data set is three times bigger than ours and is based on manual semantic annotations. We conjecture that their low gain in precision could be a side effect triggered by two factors: on the one hand, their model crucially relies on coreference chains, which are automatically generated for the test set and hence are rather noisy. On the other hand, their heuristically created training data might not represent implicit argument instances adequately.

5 Experiment 2: Implicit arguments in coherence modeling

In our second experiment, we examine the effect of implicit arguments on local coherence, i.e., the question of how well a local argument (non-)realization fits into a given context. We approach this question as follows: first, we assemble a data set of document pairs that differ only with respect to a single realization decision (Section 5.1). Given each pair in this data set, we ask human annotators to indicate their preference for the implicit or explicit argument realization in the pre-specified context (Section 5.2). Second, we attempt to emulate the decision process computationally using a discriminative model based on discourse and entity-specific features (Section 5.3).

5.1 Data compilation

We use the induced data set (henceforth *source data*), as described in Section 3, as a starting point for composing a set of document pairs that involve implicit and explicit arguments. To make sure that each document pair in this data set only differs with respect to a single realization decision, we first create two copies of each document from the source data: one copy remains in its original form, and the other copy will be modified with respect to a single argument realization. Example (2) illustrates an example of an original and modified (marked by an asterik) sentence:

(2) [The Dalai Lama's$_{A0}$] **visit** [to France$_{A1}$] ends on Tuesday.

* [The Dalai Lama's$_{A0}$] **visit** ends on Tuesday.

Note that adding and removing arguments at random can lead to structures that are semantically implausible. Hence, we restrict this procedure to predicate-argument structures (PAS) that actually occur and are aligned across two texts, and create modifications by replacing a single argument position in one text with the corresponding argument position in the comparable text. Examples (2) and (3) show two such comparable texts. The original PAS in Example (2) contains an explicit argument that is implicit in the aligned PAS and hence removed in the modified version. Vice versa, the original text in (3) involves an implicit argument, which is made explicit in the modified version.

(3) [The Dalai Lama's$_{A0}$] **visit** coincides with the Beijing Olympics.

* [The Dalai Lama's$_{A0}$] **visit** [to France$_{A1}$] coincides with the Beijing Olympics.

We ensure that the modified structure fits into the given context grammatically by only considering PAS with identical predicate form and constituent order. We found that this restriction constrains affected arguments to be modifiers, prepositional phrases and direct objects. We argue that this is actually a desirable property because more complicated alternations could affect coherence by themselves; resulting interplays would make it difficult to distinguish between the isolated effect of argument realization itself and other effects, triggered for example by sentence order (Gordon et al., 1993).

5.2 Annotation

We set up a web experiment using the NLTK package (Belz and Kow, 2011) to collect (local) coherence ratings for implicit and explicit arguments. For this experiment, we compiled a data set of 150 document pairs. As described in Section 5.1, each text pair consists of mostly the same text, with the only difference being one argument realization.

We presented all 150 pairs to two annotators[7] and asked them to indicate their preference for one alternative over the other using a continuous slider scale. The annotators got to see the full texts, with the alternatives presented next to each other. To make texts easier to read and differences easier to spot, we collapsed all identical sentences into one column and highlighted the aligned predicate (in both texts) and the affected argument (in the explicit case). An example is shown in Figure 2. To avoid any bias in the annotation process, we shuffled the sequence of text pairs and randomly assigned the side of display (left/right) of each realization type

[5] Results as reported in Tonelli and Delmonte (2011)

[6] Results computed as an average over the scores given for both test files; rounded towards the number given for the test file that contained more instances.

[7] Both annotators are undergraduate students in Computational Linguistics.

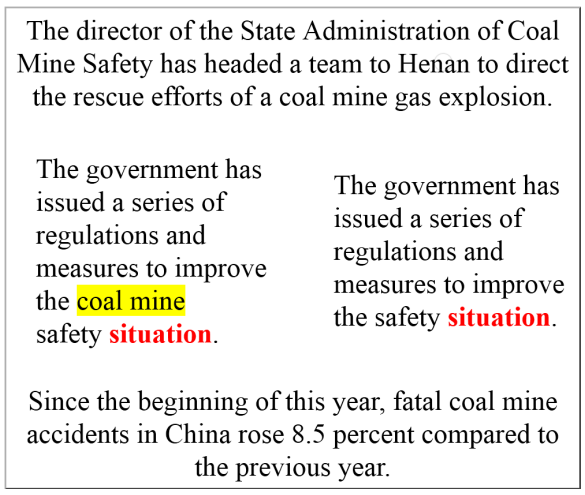

Figure 2: Texts as displayed to the annotators.

(explicit/implicit). Note that instead of providing a definition of local coherence ourselves, we simply asked the annotators to rate how "natural" a realization sounds given the discourse context.

We found that annotators made use of the full rating scale, which spans from -50 to +50, with the extremes indicating either a strong preference for the text on the left hand side or the right hand side, respectively. Most ratings are, however, concentrated more towards the center of the scale (i.e., around zero). This seems to imply that the use of implicit or explicit arguments did not make a considerable difference most of the time. The first author confirmed this assumption and resolved disagreements between annotators in several group discussions. The annotators also affirmed that some cases do not read naturally when a specific argument is or is not realized at a given position in discourse. Examples (4) and (5) illustrate two cases, in which a redundant argument is realized (A4, or *destination*) or a coherence establishing argument has been omitted (A2, or *co-signer*).[8]

(4) *?* The remaining contraband was picked up at Le Havre. The containers had **arrived** [in Le Havre] from China.

(5) *?* Lt.-Gen. Mohamed Lamari (...) denied his country wanted South African weapons to fight Muslim rebels fighting the government. "We are not going to fight a flea with a hammer," Lamari told reporters after **signing** the agreement of intent [∅].

Following discussions with the annotators, we discarded all items from the final data set, for which no clear preference could be established (72%) or the annotators had different preferences (9%). We mapped all remaining items into two classes according to whether the affected argument had to be implicit (9 texts) or explicit (20 texts). All 29 uniquely classified texts are used as a small gold standard test set for evaluation.

5.3 Coherence model

We model the decision process that underlies the (non-)realization of arguments using a SVM classifier and a range of discourse features. The features can be classified into three groups: features specific to the affected predicate-argument structure (**Parg**), the (automatic) coreference chain of the affected argument (**Coref**), and the discourse context (**Disc**).

Parg includes the absolute and relative number of realized arguments; the number of modifiers in the PAS; and the total length (in words) of the PAS and the complete sentence.

Coref includes the number of previous/follow-up mentions in a fixed sentence window; the distance (in number of words/sentences) to the previous/next mention; the distribution of occurrences over the previous/succeeding two sentences;[9] and the POS of previous/follow-up mentions.

Disc includes the total number of coreference chains in the text; the occurrence of pronouns in the current sentence; lexical repetitions in the previous/follow-up sentence; the current position in discourse (begin, middle, end); and a feature indicating whether the affected argument occured in the first sentence.

Note that most of these features overlap with those successfully applied in previous work. For example, Pitler and Nenkova (2008) also use text

[8] Note that both examples are only excerpts from the affected texts. The annotators got to see the full context.

[9] This type of feature is very similar to the transition patterns in the original entity grid. The only difference is that our features are not typed with respect to the grammatical function of explicit realizations. The reason for skipping this information lies in the insignificant amount of relevant samples in our (noisy) training data.

length, sentence-to-sentence transitions, word overlap and pronoun occurrences as features for predicting readability. Our own contribution lies in the definition of PAS-specific features and the adaptation of all features to the task of predicting (non-)realization of arguments in a predicate-argument structure.

5.4 Training data

We do not make use of any manually annotated data for training. Instead, our model relies solely on the automatically induced source data, described in Section 3, for learning. We prepare this data set as follows: first, we remove all data points that also occur in the test set. Second, we split all pairs of texts into two groups – texts that contain a predicate-argument structure in which an implicit argument has been identified (IA), and their comparable counterparts that contain the aligned PAS with an explicit argument (EA). All texts are labelled according to their group. For all texts in group EA, we remove the explicit argument from the aligned PAS. This way, the feature extractor always gets to see the text and automatic annotations as if the realization decision had not been performed and can thus extract unbiased feature values for the affected entity and argument position.

5.5 Evaluation setting

The goal of this task is to correctly predict the realization type (implicit or explicit) of an argument that maximizes the coherence of the document. As a proxy for coherence, we use the naturalness ratings given by our annotators. We evaluate classification performance on the part of our test set for which clear preferences have been established. We report results in terms of precision, recall and F_1 score. We compute precision as the fraction of correct classifier decisions divided by the total number of classifications; and recall as the fraction of correct classifier decisions divided by the total number of test items. Note that precision and recall are identical when the model provides a class label for every test item. We compute F_1 as the harmonic mean between precision and recall.

For comparison with previous work, we further apply a couple of previously proposed local coherence models: the original entity grid model by Barzilay and Lapata (2005), a modified version that uses topic models (Elsner and Charniak, 2011a) and an extended version that includes entity-specific features (Elsner and Charniak, 2011b). We further apply the discourse-new model by Elsner and Charniak (2008) and the pronoun-based model by Charniak and Elsner (2009). For all of the aforementioned models, we use their respective implementation provided with the Brown Coherence Toolkit[10]. Note that the toolkit only returns one coherence score for each document. To use the toolkit for argument classification, we use two documents per data point – one that contains the affected argument explicitly and one that does not (implicit argument) – and treat the higher scoring variant as classification output. If both documents achieve the same score, we neither count the test item as correctly nor as incorrectly classified. In contrast, we apply our own model only on the document that contains the implicit argument, and use the classifier to predict whether this realization type fits into the given context or not. Note that our model has an advantage here because it is specifically designed for this task. Yet, all models compute local coherence ratings based on entity occurrences and should thus be able to predict which realization type coheres best with the given discourse context.[11]

5.6 Results

The results are summarized in Table 4. As all models provided class labels for almost all test instances, we focus our discussion on F_1 scores. The majority class in our test set is the explicit realization type, making up 20 of the 29 test items (69%).

The original entity grid model produced differing scores for the two realization types only in 26 cases. The model exhibits a strong preference for the implicit realization type: it predicts this class in 22 cases, resulting in an F_1 score of only 15%. Taking a closer look at the features of the model reveals that this an expected outcome: in its original setting, the entity grid learns realization patterns in the form of sentence-to-sentence transitions. Most entities are, however, only mentioned a few times in a

[10] cf. `http://www.ling.ohio-state.edu/%7Emelsner/`

[11] Recall that input document pairs are identical except for the affected argument position. Consequently, the resulting coherence scores only differ with respect to affected entity realizations.

	P	R	F
Entity grid models	–	–	–
Baseline entity grid	0.15**	0.14**	0.15**
Extended entity grid	0.19**	0.17**	0.18**
Topical entity grid	0.34**	0.34**	0.34**
Other models	–	–	–
Pronouns	0.43**	0.34**	0.38**
Discourse-newness	0.48**	0.48**	0.48**
This paper	–	–	–
Our (full) model	**0.90**	**0.90**	**0.90**
Simplified model	0.83	0.83	0.83
Majority class	0.69*	0.69*	0.69*

Table 4: Results in terms of precision (P), recall (R) and F_1 score for correctly predicting argument realization; results that significantly differ from our (full) model are marked with asterisks (* $p<0.1$; ** $p<0.01$)

text, which means that non-realizations make up the 'most probable' class – independently of whether they are relevant in a given context or not. The models by Charniak and Elsner (2009) and Elsner and Charniak (2011a), which are not based on an entity grid, do not suffer from this effect and achieve better results, with F_1 scores of 38% and 48%, respectively. The topical and entity-specific refinements to the entity grid model also alleviate the bias towards non-realizations, resulting in improved F_1 scores of 18% and 34%, respectively.

To counter-balance this issue altogether, we train a simplified version of our own model that only uses features that involve occurrence patterns. The main difference between this simplified model and the original entity grid model lies in the different use of training data: while entity grid models treat all non-realized items equally, our model gets to "see" actual examples of entities that are implicit. In other words, our simplified model takes into account implicit mentions of entities, not only explicit ones. The results show that this extra information has a significant ($p<0.01$, using a randomization test (Yeh, 2000)) impact on test set performance, basically raising F_1 from 15% to 83%. Using all features of our model further increases F_1 score to 90%, the highest score achieved overall.

The highest weighted features in our model include all three feature groups: for example, the number of coreferent mentions within the preceeding/following two sentences (**Coref**), the number of words already realized in the affected predicate-argument structure (**Parg**), and the total number of coreference chains in the document (**Disc**).

6 Conclusions

In this paper, we presented a novel approach to accurately induce implicit arguments and discourse antecedents from comparable texts (cf. Section 3). We demonstrated the benefit of this kind of data for linking implicit arguments and modeling local coherence. Our experiments revealed three particularly interesting results.

Firstly, a small data set of (automatically induced) implicit arguments can have a greater impact on argument linking models than a bigger data set of artificially created instances (cf. Section 4). Secondly, the use of implicit vs. explicit arguments, while being a subtle difference in most contexts, can have a clear impact on text ratings. Thirdly, our automatically created training data enables models to learn features that considerably improve prediction of locally coherent argument realizations (cf. Section 5).

For the task of implicit argument linking, more training data will be needed to further advance the state-of-the-art. Our method for inducing this kind of data, by exploiting aligned predicate-argument structures from comparable texts, has shown promising results. Future work will have to explore this direction more fully, for example, by identifying ways to induce data with higher recall. Integrating argument (non-)realization into a full model of local coherence also remains part of future work. In this paper, we presented a suitable basis for such work: a training set that contains empirical data on implicit arguments in discourse; and a feature set that models argument realization with high accuracy.

Acknowledgments

We are grateful to the Landesgraduiertenförderung Baden-Württemberg for funding within the research initiative "Coherence in language processing" at Heidelberg University. We thank our annotators and four anonymous reviewers.

References

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* Ann Arbor, Michigan, USA, 25–30 June 2005, pages 141–148.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Anja Belz and Eric Kow. 2011. Discrete vs. continuous rating scales for language evaluation in nlp. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 230–235, Portland, Oregon, USA, June.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstration Volume*, pages 33–36, Beijing, China, August. Coling 2010 Organizing Committee.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China, August.

Cheryl Brown. 1983. Topic continuity in written english narrative. In Talmy Givon, editor, *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. John Benjamins, Amsterdam, The Netherlands.

Eugene Charniak and Micha Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 148–156, Athens, Greece, March.

Desai Chen, Nathan Schneider, Dipanjan Das, and Noah A. Smith. 2010. SEMAFOR: Frame argument resolution with log-linear models. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 264–267, Uppsala, Sweden, July.

Micha Elsner and Eugene Charniak. 2008. Coreference-inspired coherence modeling. In *Proceedings of ACL-08: HLT, Short Papers*, pages 41–44, Columbus, Ohio, June.

Micha Elsner and Eugene Charniak. 2011a. Disentangling chat with local coherence models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1179–1189, Portland, Oregon, USA, June.

Micha Elsner and Eugene Charniak. 2011b. Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 125–129, Portland, Oregon, USA, June.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts, USA.

Katja Filippova and Michael Strube. 2007. Extending the entity-grid coherence model to semantically related entities. In *Proceedings of the 11th European Workshop on Natural Language Generation,* Schloss Dagstuhl, Germany, 17–20 June 2007, pages 139–142.

C. J. Fillmore. 1986. Pragmatically controlled zero anaphora. In *Proceedings of the twelfth annual meeting of the Berkeley Linguistics Society*, pages 95–107.

Matthew Gerber and Joyce Chai. 2012. Semantic Role Labeling of Implicit Arguments for Nominal Predicates. *Computational Linguistics*, 38(4):755–798.

Peter C. Gordon, Barbara J. Grosz, and Laura A. Gilliom. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17:311–347.

Philip Gorinski, Josef Ruppenhofer, and Caroline Sporleder. 2013. Towards weakly supervised resolution of null instantiations. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 119–130, Potsdam, Germany, March.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Aravind K. Joshi and Steve Kuhn. 1979. Centered logic: The role of entity centered sentence representation in natural language inferencing. In *Proceedings of the 6th International Joint Conference on Artificial Intelligence,* Tokyo, Japan, August, pages 435–439.

Egoitz Laparra and German Rigau. 2012. Exploiting explicit annotations and semantic types for implicit argument resolution. In *Proceedings of the Sixth IEEE International Conference on Semantic Computing (ICSC 2010)*, pages 75–78, Palermo, Italy, September. IEEE Computer Society.

Egoitz Laparra and German Rigau. 2013. Sources of evidence for implicit argument resolution. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 155–166, Potsdam, Germany, March.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500, Jeju Island, Korea, July.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013.

Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4). Accepted for publication.

Annie Louis and Ani Nenkova. 2010. Creating local coherence: An empirical assessment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 313–316, Los Angeles, California, June.

Sebastian Martschat, Jie Cai, Samuel Broscheit, Éva Mújdricza-Maydt, and Michael Strube. 2012. A multigraph model for coreference resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 100–106, Jeju Island, Korea, July.

Neil McIntyre and Mirella Lapata. 2009. Learning to tell tales: A data-driven approach to story generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing,* Singapore, 2–7 August 2009, pages 217–225.

Adam Meyers, Ruth Reeves, and Catherine Macleod. 2008. *NomBank v1.0*. Linguistic Data Consortium, Philadelphia.

Tatjana Moor, Michael Roth, and Anette Frank. 2013. Predicate-specific annotations for implicit role binding: Corpus annotation, data analysis and evaluation experiments. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*, pages 369–375, Potsdam, Germany, March.

Martha S. Palmer, Deborah A. Dahl, Rebecca J. Schiffman, Lynette Hirschman, Marcia Linebarger, and John Dowding. 1986. Recovering implicit information. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics,* New York, N.Y., 10–13 June 1986, pages 10–19.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Robert Parker, David Graff, Jumbo Kong, Ke Chen, and Kazuaki Maeda. 2011. *English Gigaword Fifth Edition*. Linguistic Data Consortium, Philadelphia.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii, October.

Michael Roth and Anette Frank. 2012a. Aligning predicate argument structures in monolingual comparable texts: A new corpus for a new task. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 218–227, Montreal, Canada, June.

Michael Roth and Anette Frank. 2012b. Aligning predicates across monolingual comparable texts using graph-based clustering. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 171–182, Jeju Island, Korea, July.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2010. SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluations*, pages 45–50, Uppsala, Sweden, July.

Ivan A. Sag and Jorge Hankamer. 1984. Towards a Theory of Anaphoric Processing. *Linguistics and Philosophy*, 7:325–345.

Candace L. Sidner. 1979. Towards a computational theory of definite anaphora comprehension in English. Technical Report AI-Memo 537, Massachusetts Institute of Technology, AI Lab, Cambridge, Mass.

Carina Silberer and Anette Frank. 2012. Casting implicit role linking as an anaphora resolution task. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*, pages 1–10, Montréal, Canada, 7-8 June.

Michael K. Tanenhaus and Greg N. Carlson. 1990. Comprehension of Deep and Surface Verbphrase Anaphors. *Language and Cognitive Processes*, 5(4):257–280.

Sara Tonelli and Rodolfo Delmonte. 2010. VENSES++: Adapting a deep semantic processing system to the identification of null instantiations. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 296–299, Uppsala, Sweden, July.

Sara Tonelli and Rodolfo Delmonte. 2011. Desperately seeking implicit arguments in text. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, pages 54–62, Portland, Oregon, USA, June.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2011. *OntoNotes Release 4.0*. Linguistic Data Consortium, Philadelphia.

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

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 947–953, Saarbrücken, Germany, August.