Inducing Implicit Arguments from Comparable Texts: A Framework and Its Applications

Michael Roth*
University of Edinburgh

Anette Frank**
Heidelberg University

In this article, we investigate aspects of sentential meaning that are not expressed in local predicate–argument structures. In particular, we examine instances of semantic arguments that are only inferable from discourse context. The goal of this work is to automatically acquire and process such instances, which we also refer to as implicit arguments, to improve computational models of language. As contributions towards this goal, we establish an effective framework for the difficult task of inducing implicit arguments and their antecedents in discourse and empirically demonstrate the importance of modeling this phenomenon in discourse-level tasks.

Our framework builds upon a novel projection approach that allows for the accurate detection of implicit arguments by aligning and comparing predicate–argument structures across pairs of comparable texts. As part of this framework, we develop a graph-based model for predicate alignment that significantly outperforms previous approaches. Based on such alignments, we show that implicit argument instances can be automatically induced and applied to improve a current model of linking implicit arguments in discourse. We further validate that decisions on argument realization, although being a subtle phenomenon most of the time, can considerably affect the perceived coherence of a text. Our experiments reveal that previous models of coherence are not able to predict this impact. Consequently, we develop a novel coherence model, which learns to accurately predict argument realization based on automatically aligned pairs of implicit and explicit arguments.

1. Introduction

The goal of semantic parsing is to automatically process natural language text and map the underlying meaning of text to appropriate representations. **Semantic role labeling**

* School of Informatics, University of Edinburgh, EH8 9AB Edinburgh, United Kingdom.
E-mail: mroth@inf.ed.ac.uk.
** Department of Computational Linguistics, Heidelberg University, 69120 Heidelberg, Germany.
E-mail: frank@cl.uni-heidelberg.de.

Submission received: 5 June 2014; revised version received: 13 April 2015; accepted for publication: 28 August 2015.

doi:10.1162/COLI_a_00236

induces shallow semantic representations, so-called predicate–argument structures, by processing sentences and mapping them to predicates and associated arguments. Arguments of these structures can, however, be non-local in natural language text, as shown in Example 1.

Example 1

(a) El Salvador is the only Latin American country which has troops in *Iraq*.

(b) Nicaragua withdrew its troops last month.[1]

Applying a semantic role labeling system on sentence (1b) produces a representation that consists of the predicate *withdraw*, a temporal modifier (*last month*) and two associated arguments: the entity withdrawing (*Nicaragua*) and the thing being withdrawn (*its troops*). From the previous sentence (1a), we can additionally infer a third argument: namely, the source from which Nicaragua withdrew its troops (*Iraq*). By leaving this piece of information implicit, the text fragment in sentence (1b) illustrates a typical case of non-local, or implicit, role realization (Gerber and Chai 2012). In this article, we view implicit arguments as a discourse-level phenomenon and treat corresponding instances as implicit references to discourse entities.

Taking this perspective, we build upon previous work on discourse analysis. Following Sidner (1979) and Joshi and Kuhn (1979), utterances in discourse typically focus on a set of *salient* entities, which are also called the *foci* or *centers*. Using the notion of centers, Grosz, Joshi, and Weinstein (1995) defined the Centering framework, which relates the salience of an entity in discourse to linguistic factors such as choice of referring expression and syntactic form.[2] Both extremes of salience, that is, contexts of referential continuity and irrelevance, can also be reflected by the *non-realization* of an entity (Brown 1983). Specific instances of this phenomenon, so-called zero anaphora, have been well-studied in pro-drop languages such as Japanese (Kameyama 1985), Turkish (Turan 1995), and Italian (Di Eugenio 1990). For English, only a few studies exist that explicitly investigated the effect of non-realizations on coherence. Existing work suggests, however, that indirect references and non-realizations are important for modeling and measuring coherence (Poesio et al. 2004; Karamanis et al. 2009), respectively, and that such phenomena need to be taken into consideration to explain local coherence where adjacent sentences are neither connected by discourse relations nor in terms of coreference (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 illustrated a text fragment, in which argument realization is necessary in the first sentence but redundant in the second. That is, mentioning *Iraq* in the second sentence is not necessary (for a human being) to understand the meaning of the text. In contrast, making both references explicit, as shown in Example 2, would be redundant and could lead to the perception that the text is merely a concatenation of two independent sentences — rather than a set of adjacent sentences that form a meaningful, or *coherent*, discourse.

1 In the remainder of this article, we use the following notation for predicate–argument structures: underlined words refer to predicates and [possibly empty] constituents in [square brackets] refer to arguments.

2 We note that the relation between coherence and choice of referring expressions has also been studied outside of the Centering framework. A comprehensive overview of research in linguistics on this subject can be found in Arnold (1998).

Example 2

(a) El Salvador is now the only Latin American country which has troops in Iraq.

(b) [Nicaragua] withdrew [its troops] [from Iraq] last month.

The phenomenon of implicit arguments has neither been extensively studied in the context of semantic role labeling nor in coherence modeling. One of the main reasons for this lies in the fact that models for these tasks are typically developed on the basis of annotated corpora. In contrast, implicit arguments are not overt in text and are difficult to uncover, hence there exist only few and small data sets that contain respective annotations explicitly. In this article, we present a novel framework for computationally inducing a corpus with automatic annotations of *implicit arguments and their respective antecedents in discourse*. To achieve this goal, our framework exploits pairs of *comparable texts*, which convey information about the same events, states, and entities but potentially differ in terms of depth and perspective. These differences can also affect argument realization, making it possible to identify instances of co-referring but only partially overlapping argument structures. Given automatically induced instances of implicit arguments and their document contexts, we show how these can be utilized to enhance current models of local coherence and implicit argument linking. The resulting models and research insights will be of importance for many applications that involve the understanding or generation of natural language text beyond the sentence level.

2. Overview

The goal of this work is to provide data and methods to better capture the phenomenon of implicit arguments in semantic role labeling and coherence modeling. To accomplish this goal, we propose inducing suitable training data automatically. For semantic role labeling, this kind of training data will contain explicit links between implicit arguments and their respective discourse antecedents; for modeling coherent argument realization, the data will also provide discourse contexts of explicit and implicit references to the same entity. We propose inducing such information by exploiting pairs of comparable texts. Our motivation for this lies in the fact that comparable texts convey by and large the same information but differ in depth and presentation. This means that while we expect two comparable texts to contain references to the same events and participants, an entity might be understood implicitly at a specific point in one text (because it can be inferred from context), whereas an explicit mention might be necessary in a comparable text (given the different context). Based on this assumption, our method aims to find complementary (explicit) information in pairs of texts, which can be aligned and subsequently be projected to identify missing (implicit) pieces in one another. Hence, we separate our task into two parts. In the first part, described in Section 4, we ask the question:

Q_1: How can we detect predicate–argument structures that are shared across two discourses?

Based on a corpus of comparable texts, we propose a new task of aligning predicate–argument structures (PASs) across complete discourses. We define this task on the basis of semantic annotations, which we compute automatically using a state-of-the-art semantic role labeler. For the alignment task itself, we develop a graph-based clustering

model. We show that by taking into account features on the levels of predicate, arguments, and discourse context, this model is able to predict accurate alignments across comparable texts, without relying on preprocessing methods that identify parallel text fragments. Based on automatically aligned structures, the second part of our approach aims to identify and resolve implicit arguments by projecting explicit information from one text to another. The question for this part of our approach (cf. Section 5) is:

Q_2: How can we find implicit arguments and their antecedents in discourse?

Our approach to answering this question relies solely on information that can be computed automatically: We first identify potential instances of implicit arguments by comparing two aligned predicate–argument structures; for each entity that is mentioned in one PAS (explicit argument) but not in an aligned structure (implicit argument), we then apply a cross-document coreference resolution technique to also find co-referring entity mentions in the document in which an implicit reference was found.

Finally, we demonstrate the utility of automatically induced instances of implicit arguments in two task-based settings: linking arguments in discourse and modeling local coherence. The question here is:

Q_3: How can induced argument instances and their respective discourse contexts be utilized to enhance NLP models?

We address both tasks by applying our automatically induced data set as training material to improve the performance of supervised models. For the first task, described in Section 6, we apply our data to enhance training of an existing system that identifies and links implicit arguments in discourse. To evaluate the impact of our induced data set, we test the modified model on a standard evaluation data set, on which we compare our results with those of previous work. For the second task, we develop a new coherence model that predicts whether an argument realization or non-realization in context improves the perceived coherence of the affected segment in discourse. We evaluate this coherence model in an intrinsic evaluation scenario, in which we compare model predictions to human judgments on argument use (cf. Section 7).

3. Background

This article draws on insights from computational linguistic research on implicit arguments and coherence modeling as well as from previous work on inducing semantic resources. It is further related to recent work in paraphrasing and event coreference.

3.1 Implicit Arguments and Coherence Modeling

The goal of this work is to induce instances of implicit arguments, together with their discourse antecedents, and to utilize them in semantic processing and coherence modeling. This section summarizes previous work on implicit arguments and coherence modeling, and provides an outlook on how instances of implicit arguments can be of use in a novel entity-based model of local coherence.

Implicit Arguments. The role of implicit arguments was studied early on in the context of semantic processing (Fillmore 1986; Palmer et al. 1986), although most semantic

role labeling systems nowadays operate solely within local syntactic structures and do not perform any additional inference regarding missing information. First data sets that focus 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 NomBank (Meyers, Reeves, and Macleod 2008), and Moor, Roth, and Frank (2013) provided annotations for parts of the OntoNotes corpus (Weischedel et al. 2011). All three resources are, however, severely limited: Annotations in the latter two studies are restricted to 10 and 5 predicate types, respectively; the training set of the SemEval task, in contrast, consists of full-text annotations for all occurring predicates but contains only 245 instances of resolved implicit arguments in total. All groups working on the shared task identified data sparsity as one of the main issues (Chen et al. 2010; Ruppenhofer et al. 2012; Laparra and Rigau 2013). Silberer and Frank (2012) point out that additional training data can be heuristically created by treating anaphoric pronoun mentions as implicit arguments. Their experimental results confirmed 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. As observed in related work on classifying discourse relations, information learned from artificial training data might not always generalize well to naturally occurring examples (Sporleder and Lascarides 2008). To automatically create data that is linguistically more similar to manually labeled implicit arguments, we introduce an alternative method that induces instances of implicit arguments from a raw corpus of comparable texts.

Coherence Modeling. In the context of coherence modeling, much previous work has focused on entity-based approaches, with the most prominent model being the entity grid by Barzilay and Lapata (2005). This model has originally been proposed for automatic sentence ordering but has since also been applied in coherence evaluation and readability assessment (Barzilay and Lapata 2008; Pitler and Nenkova 2008), story generation (McIntyre and Lapata 2010), and authorship attribution (Feng and Hirst 2014). Based on the original model, several 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, solely based on explicit entity mentions, resulting in insufficient representations when dealing with inferable references. Example 3 illustrates this shortcoming.

Example 3

	implicit roles
(a) 27 tons of cigarettes were picked up in Le Havre.	`agent`
(b) The containers had arrived yesterday.	`content, destination`

The two sentences in the given example do not share any explicit references to the same entity. Hence, none of the aforementioned models is able to correctly predict that both sentences cohere in the presented order—but not in the reversed order.

As discussed in more detail by Poesio et al. (2004), such cases can still be resolved when taking into account *indirect* realizations, that is, associative references to a discourse entity. Accordingly, recent work by Hou, Markert, and Strube (2013) proposed

considering bridging resolution to fill specific gaps in the entity grid model. Following this argument, *the containers* in sentence (3b) could be understood as a bridging anaphor that refers back to the *27 tons of cigarettes* in sentence (3a). As mentioned in the NomBank annotation guidelines (Meyers 2007), some cases of nominal predicates and their arguments do coincide with bridging relations and we treat them as implicit arguments in this article. For example, when interpreting *container* as a nominalization of the verbal predicate CONTAIN, the *cigarettes* from the previous sentence can be understood as one of its arguments: namely, its content. Not all cases of implicit arguments involve bridging anaphora and vice versa, however. The main difference lies in the fact that bridging typically describes a relation between entities. Hence, bridging instances only coincide with those of implicit arguments if the bridging anaphor is linguistically realized as a (nominal) predicate (*container* in Example 3b). In contrast, implicit arguments frequently form parts of predicate–argument structures that refer to events, states, and entities that are not bridging anaphora. For example, the destination of the predicate ARRIVE in sentence (3b) can also be interpreted as an implicit argument, namely, *Le Havre* from the preceding context.

Taking implicit arguments into consideration, we can see that the second sentence refers back to two previously introduced entities, *Le Havre* and *cigarettes*, reflecting the fact that the sentences cohere in the presented order. In Section 7, we present a novel model of local coherence that aims to capture the effect of argument realizations on perceived coherence. As an overall goal, this model should be able to predict whether an entity reference should be realized explicitly to establish (local) coherence—or whether the entity can already be understood from context. Based on such predictions, the model can be applied in text generation to ensure that necessary references are explicit and that redundant repetitions are avoided.

3.2 Semantic Resource Induction

The methods applied in this article are based on ideas from previous work on inducing semantic resources from parallel and comparable texts. Most work in this direction has been done in the context of cross-lingual settings, including the learning of translations of words and phrases using statistical word alignments (Kay and Röscheisen 1993; DeNero, Bouchard-Côté, and Klein 2008, inter alia) and approaches to projecting annotations from one language to another (Yarowsky and Ngai 2001; Kozhevnikov and Titov 2013, inter alia). In the following, we discuss previous approaches to annotation projection as well as related work in paraphrasing and event coreference.

Annotation Projection. A widespread method for the induction of semantic resources is the so-called annotation projection approach. The rationale of this approach is to induce annotated data in one language, given already-annotated instances in another language. As an example, semantic role annotations of a text in English can be transferred to a parallel text in order to induce annotated instances for a lexicon in another language (Padó and Lapata 2009). In previous work, this method has been applied on various levels of linguistic analysis: from syntactic information in the form of part-of-speech tags and dependencies (Yarowsky and Ngai 2001; Hwa et al. 2005), through annotation of temporal expressions and semantic roles (Spreyer and Frank 2008; van der Plas, Merlo, and Henderson 2011), to discourse-level phenomena such as coreference (Postolache, Cristea, and Orasan 2006). All of the aforementioned instances of the projection approach make use of the same underlying technique: Firstly, words are

aligned in a parallel corpus using statistical word alignment; secondly, annotations on a single word or between multiple words in one text are transferred to the corresponding aligned word(s) in the parallel text. This procedure typically assumes that two parallel sentences express the same meaning. A notable exception is the work by Fürstenau and Lapata (2012), which utilizes alignments between syntactic structures to "project" semantic role information from a role-annotated corpus to unseen sentences that are selected from a corpus in the same language.

In our work, we apply annotation projection to monolingual comparable texts. In comparison to parallel texts, we have to account for potential differences in perspective and detail that make our task—in particular, the alignment sub-task—considerably more difficult. In contrast to Fürstenau and Lapata's setting, which involves incomparable texts, we assume that text pairs in our setting still convey information on the same events and participants. This means that in addition to aligning predicate–structures across texts, we can also merge complementary details realized in each structure. We achieve this by making use of the same principle that underlies the projection approach: Given annotation that is available in one text (explicit argument), we project this information to a text in which the corresponding annotation is not available (implicit argument). Because our task is based in a monolingual setting, we can make use of the same preprocessing tools across texts.

Paraphrasing and Event Coreference. We overcome the difficulty of inducing word alignments across comparable texts by computing alignments on the basis of predicate–argument structures. Using predicate–argument structures as targets makes our setting related to previous work on paraphrase detection and coreference resolution of event mentions. Each of these tasks focuses, however, on a different level of linguistic analysis from ours: Following the definitions embraced by Recasens and Vila (2010), "paraphrasing" is a relation between two lexical units that have the same meaning, whereas "coreference" indicates that two referential expressions point to the same referent in discourse.[3] In contrast to work on paraphrasing, we are specifically interested in pairs of text fragments that involve implicit arguments, which can only be resolved in context.

In line with our goal of inducing implicit arguments, we define the units or expressions to be aligned in our task as the predicate–argument structures that can (automatically) be identified in text. This task definition further makes our task distinct from event coreference, where coreference is established based on a pre-specified set of events, reference types, or definitions of event identity (Walker et al. 2006; Pradhan et al. 2007; Lee et al. 2012, inter alia). Although corresponding annotations can certainly overlap with those in our task, we emphasize that the focus of our work is not to find all occurrences of co-referring events. Instead, our goal is to align predicate–argument structures that have a common meaning in context across different discourses. Hence, we neither consider intra-document coreference nor pronominal event references here. As alignable units in our work are not restricted to a pre-specified definition of event or event identity, the task addressed here involves any kind of event, state, or entity that is linguistically realized as a predicate–argument structure. Examples that go beyond traditional event coreference include in particular noun phrases such as "a [rival] of the company" and "the [finance] director [of IBM]"(cf. Meyers, Reeves, and Macleod 2008).

3 Note that in the remainder of this article, we use the term "coreference" in a wider sense to also encompass referents to events and entities in the real world, rather than just referents grounded in discourse.

4. Induction Framework Step 1: Aligning Predicate–Argument Structures

In the first step of our induction framework, we align predicate–argument structures (PASs) across pairs of comparable text. To execute this task properly, we define annotation guidelines and construct a gold standard for the development and evaluation of alignment models. All annotations are performed on pairs of documents from a corpus of comparable texts. We introduce this corpus and describe our annotations in Section 4.1. We describe a graph-based model that we developed for the automatic alignment in Section 4.2. Finally, we present an experimental evaluation of our model, together with another recently proposed model and several baselines, in Section 4.3.

4.1 Corpus and Annotation

As a basis for aligning predicate–argument structures across texts, we make use of a data set of comparable texts extracted from the English Gigaword corpus (Parker et al. 2011). The Gigaword corpus is one of the largest English corpora available in the news domain and contains over 9.8 million articles from seven newswire agencies that report on (the same) real-world incidents. The data set of comparable texts used in this work contains 167,728 pairs of articles that were extracted by matching the headlines of texts published within the same time frame (Roth and Frank 2012a). A set of such document headlines is given in Example 4:

Example 4
India fires tested anti-ship cruise missile
(*Xinhua News Agency*, 29 October 2003)

India tests supersonic cruise anti-ship missile
(*Agence France Presse*, 29 October 2003)

URGENT: India tests anti-ship cruise missile
(*Associated Press Worldstream*, 29 October 2003)

We preprocess each article in the set of 167,728 pairs using the MATE tools (Björkelund et al. 2010; Bohnet 2010), including a state-of-the-art semantic role labeler that identifies PropBank/NomBank-style predicate–argument structures (Palmer, Gildea, and Kingsbury 2005; Meyers, Reeves, and Macleod 2008). Based on the acquired PAS, we perform manual alignments. In Section 4.1.1, we summarize the annotation guidelines for this step. An overview of the resulting development and evaluation data set is provided in Section 4.1.2.

4.1.1 Manual Annotation. We selected 70 pairs of comparable texts and asked two annotators to manually align predicate–argument structures obtained from preprocessing. Both annotators were students in computational linguistics, one undergraduate and one postgraduate. The texts were selected with the constraint that each text consists of 100 to 300 words. We chose this constraint as longer text pairs seemed to contain a higher number of unrelated predicates, making the alignment tasks difficult to manage for the annotators. Both annotators received detailed guidelines that describe alignment requirements and the overall procedure.[4] We summarize essentials in the following.

4 cf. http://projects.cl.uni-heidelberg.de/india/files/guidelines.pdf.

Sure and Possible Links. Following standard practice in word alignment (cf. Cohn, Callison-Burch and Lapata 2008, inter alia) the annotators were instructed to distinguish between *sure* (S) and *possible* (P) alignments, depending on how certainly, in their opinion, two predicates (including their arguments) describe the same event, state, or entity. Examples 5 and 6 show a sure and possible predicate pairing, respectively.

Example 5
The regulator ruled on September 27 that Nasdaq was qualified to bid.
The authority had already approved a similar application by Nasdaq.

Example 6
Myanmar's government said it has released some 220 political prisoners.
The government has been regularly releasing members of Suu Kyi's party.

Replaceability. As a rule of thumb for deciding whether to align two structures, annotators were told to check how well the affected predicate–argument structures could be replaced by one another in their given context.

Missing Context. In case one text does not provide enough context to decide whether two predicates in the paired documents refer to the same event, an alignment should not be marked as sure.

Similar Predicates. Annotators were told explicitly that sure links can be used even if two predicates are semantically different but have the same meaning in context. Example 7 illustrates such a case.

Example 7
The volcano roared back to life two weeks ago.
It began erupting last month.

1-to-1 vs. n-to-m. We asked the annotators to find as many 1-to-1 correspondences as possible and to prefer 1-to-1 matches over *n*-to-*m* alignments. In case of multiple mentions of the same event, we further asked the annotators to provide only one sure link per predicate and mark remaining cases as possible links. As an additional guideline, annotators were asked to only label the PAS pair with the highest information overlap as a sure link. If there is no difference in information overlap, the predicate pair that occurs first in both texts should be marked as a sure alignment. The intuition behind this guideline is that the first mention introduces the actual event whereas later mentions just (co-)refer or add further information.

4.1.2 Resulting Data Set. In total, the annotators (A/B) aligned 487/451 sure and 221/180 possible alignments. Following Brockett (2007), we computed agreement on labeled annotations, including unaligned predicate pairs as an additional *null* category. We computed κ following Fleiss, Levin, and Paik (1981) and observed an overall score of 0.62, with κ values per category of 0.74 and 0.19 for *sure* and *possible* alignments, respectively. The numbers show that both annotators substantially agree on which pairs of predicate–argument structures "surely" express the same proposition. Identifying further references to the same event or state, in contrast, can only be achieved with

Table 1
Statistics on predicates and alignments in the annotated data sets.

	Development	Evaluation
number of text pairs	10	60
number of preprocessed predicates		
all predicates (average)	395 (39.5)	3,453 (57.5)
nouns only (average)	168 (16.8)	1,531 (25.5)
verbs only (average)	227 (22.7)	1,922 (32.0)
number of alignments		
all alignments (average)	78 (7.8)	807 (13.4)
sure only (average)	35 (3.5)	446 (7.4)
possible only (average)	43 (4.3)	361 (6.0)
properties of aligned PASs		
same POS (nouns/verbs)	88.5% (24/42)	82.4% (242/423)
same lemma (total)	53.8% (42)	47.5% (383)
unequal number of arguments (total)	30.8% (24)	39.7% (320)

fairly low agreement. For the construction of a gold standard, we take the intersection of all *sure* alignments by both annotators and the union of all *possible* alignments.[5] We further resolved cases that involved a sure alignment on which the annotators disagreed in a group discussion and added them to our gold standard accordingly. We split the final corpus into a development set of 10 document pairs and an evaluation set of 60 document pairs.

Table 1 summarizes information about the resulting annotations in the development and evaluation set. As can be seen, the documents in the development set contain a smaller number of predicates (39.5 vs. 57.6) and alignments (8.7 vs. 13.4) on average. The fraction of aligned predicates is, however, about the same (22.0% vs. 23.3%). Across both data sets, the average numbers of observed predicates is approximately 55, of which 31 are verbs and 24 are nouns. In the development and evaluation sets, the average number of sure alignments are 3.5 and 7.4. From all aligned predicate pairs in both data sets, 82.6% are the same part of speech (30.0% both nouns, 52.6% both verbs). In total, 48.0% of all alignments are between predicates of identical lemmata. As a rough indicator for diverging argument structures captured in the annotated alignments, we analyzed the number of aligned predicates that involve a different number of realized arguments. In both data sets together, this criterion applied in 344 cases (38.9% of all alignments).

4.2 Alignment Model

For the automatic alignment of predicate–argument structure alignments across texts, we opt for an unsupervised graph-based method. That is, we represent pairs of documents as bipartite graphs and subsequently aim to separate a graph into subgraphs that represent corresponding predicate–argument structures across texts. We define our general graph representation for this task in Section 4.2.1. In Section 4.2.2, we introduce

5 In our evaluation, only *sure* alignments need to be predicted by a system, whereas *possible* alignments are optional and not counted towards the attested recall (cf. Section 4.3).

a range of similarity measures that are used to weight edges in the graph representation. In Section 4.2.3, we introduce our algorithm to separate graphs, representing pairs of documents, into subgraphs of corresponding predicates. The overall model is an extended variant of the clustering approach described in Roth and Frank (2012b) and uses an enhanced set of similarity measures.

4.2.1 Graph Representation. We build a bipartite graph representation for each pair of texts, using as vertices the predicate–argument structures assigned during preprocessing (cf. Section 4.1). We represent each predicate as a node and integrate information about arguments implicitly. Given the sets of predicates P_1 and P_2 of two comparable texts T_1 and T_2, respectively, we formally define an undirected graph G_{P_1,P_2} following Equation 7.

$$G_{P_1,P_2} = \langle V, E \rangle \quad \text{where } \begin{matrix} V = P_1 \cup P_2 \\ E = P_1 \times P_2 \end{matrix} \tag{7}$$

Edge Weights. We specify the edge weight between two nodes representing predicates $p_1 \in P_1$ and $p_2 \in P_2$ as a weighted linear combination of the similarity measures S described in the next section.

$$w_{p_1 p_2} = \sum_{i}^{|S|} \lambda_i * \text{sim}_i(p_1, p_2) \tag{8}$$

Initially we set all weighting parameters λ_i to have uniform weights by default. We describe a tuning routine to find an optimized weighting scheme for the individual measures in the experimental evaluation of our approach (Section 4.3).

4.2.2 Similarity Measures. We use a number of similarity measures that make use of complementary information on the predicates, arguments, and discourse context of two predicate–argument structures. Given two lemmatized predicates p_1, p_2 and their sets of arguments $A_1 = \text{args}(p_1)$, $A_2 = \text{args}(p_2)$, we define seven measures in total. Three of them are specific to the predicates themselves, two take into account information on associated arguments, and two measures capture discourse-level properties. The predicate-specific measures as well as one argument-specific measure correspond to measures previously described in Roth and Frank (2012b). This article extends previous work by considering discourse-level information and an additional argument-specific measure that takes into account argument labels. We demonstrate the benefits of these measures in practice in Section 4.3.

Similarity in WordNet. Given all synsets that contain the two predicates p_1, p_2, we compute their similarity in WordNet (Fellbaum 1998) as the maximal pairwise score calculated using the information content based measure proposed by Lin (1998). We rely on the WordNet hierarchy to find the least common subsumer (lcs) of two synsets and use the pre-computed Information Content (IC) files from Pedersen (2010) to compute this measure as defined in Equation (9).

$$\text{sim}_{\text{WN}}(p_1, p_2) = \max_{\langle s_1, s_2 \rangle : s_i \in \text{synsets}(p_i)} \frac{\text{IC}(\text{lcs}(s_1, s_2))}{\text{IC}(s_1) * \text{IC}(s_2)} \tag{9}$$

Similarity in VerbNet. We additionally make use of VerbNet (Kipper et al. 2008) to compute similarities between verb pairs that cannot be captured by WordNet relations. Verbs in VerbNet are categorized into classes according to their meaning as well as syntactic behavior. A verb class C can recursively embed sub-classes $C_s \in \text{sub}(C)$ that represent finer semantic and syntactic distinctions. In Equation (10), we define a simple similarity function that assigns fixed scores to pairs of predicates p_1, p_2 depending on their relatedness within VerbNet.[6]

$$\text{sim}_{\text{VN}}(p_1, p_2) = \begin{cases} 1.0 & \text{if } \exists C : p_1, p_2 \in C \\ 0.8 & \text{if } \exists C, C_s : C_s \in \text{sub}(C) \\ & \quad \wedge ((p_1 \in C, p_2 \in C_s) \vee (p_1 \in C_s, p_2 \in C)) \\ \textit{default} & \text{else} \end{cases} \tag{10}$$

We empirically set the *default* value to the average VerbNet similarity (with unrelated pairs counted as 0.0) computed over one million random pairs of predicates in our corpus.

Similarity in a Semantic Space. As predicates can be absent from WordNet and VerbNet, or distributed over separate hierarchies due to different parts-of-speech (verbal vs. nominal predicates), we additionally calculate similarity based on distributional meaning in a semantic space (Landauer and Dumais 1997). This measure is based on the similarity of contexts of two given predicates over all their instances in a corpus. To compute this measure, we first calculate the Pointwise Mutual Information (PMI) for each predicate $p \in \{p_1, p_2\}$ and the n most frequent context words $c \in C$ following Equation (11).

$$\text{pmi}(p, c) = \frac{\text{freq}(p, c)}{\text{freq}(p) * \text{freq}(c)} \tag{11}$$

As we are dealing with predicates of different parts-of-speech, we calculate joint frequencies in terms of context windows instead of relying on syntactic dependencies as proposed in more recent approaches to distributional semantics (Padó and Lapata 2007; Erk and Padó 2008; Baroni and Lenci 2010). More precisely, we extract context windows of five words to the left and to the right from the Gigaword corpus (Parker et al. 2011), and compute the PMI for the 2,000 most frequent context words $c_1 \ldots c_{2,000} \in C$. The same setting has been successfully applied in related tasks, including word sense disambiguation (Guo and Diab 2011) and measuring phrase similarity (Mitchell and Lapata 2010). Vector representations are computed following Equation (12), and similarities are calculated as the cosine function of the angle between two vectors, as defined in Equation (13).

$$\vec{p} = (\text{pmi}(p, c_1), \text{pmi}(p, c_2), \ldots, \text{pmi}(p, c_{2,000})) \tag{12}$$

$$\text{sim}_{\text{Dist}}(p_1, p_2) = \frac{\vec{p_1} \cdot \vec{p_2}}{||\vec{p_1}|| * ||\vec{p_2}||} \tag{13}$$

6 Note that the weight of 0.8 was set in an ad hoc manner (instead of being optimized) in order to avoid overfitting on our small development corpus.

Bag-of-Words Similarity. As a simple argument-specific measure, we compute the overlap of word tokens over all (concatenated) arguments of each predicate–argument structure (PAS). Formally, this similarity measure considers all arguments $a_1 \in A_1$ and $a_2 \in A_2$ associated with the predicates p_1 and p_2, respectively, and calculates overlap as defined in Equation (14). In order to control the impact of frequently occurring words, we weight each word by its Inverse Document Frequency (IDF), which we calculate over all documents d in our corpus D:

$$\text{sim}_{\text{ABoW}}(p_1, p_2) = \frac{\sum_{w \in A_1 \cap A_2} \text{idf}(w)}{\sum_{w \in A_1} \text{idf}(w) + \sum_{w \in A_2} \text{idf}(w)} \tag{14}$$

$$\text{idf}(w) = \log \frac{|D|}{|\{d \in D | w \in \text{words}(d)\}|} \tag{15}$$

Head of Arguments Similarity. We further define an argument-specific measure that only compares the semantic heads of arguments that have the same argument label. For example, given two PASs that consist of predicates p_1, p_2 and arguments labeled A0 and A1, we compute the similarity of the two arguments labeled A0 (also denoted as $\text{label}(a) = \text{'A0'}$) and the similarity of the two arguments labeled A1 ($\text{label}(a) = \text{'A1'}$). Each similarity between argument heads is computed using the WordNet-based measure described earlier. We extract the semantic head of each argument by considering the dependency tree, as predicted by MATE tools, in which we look for the noun or verb on the highest level within the argument span. Finally, we collapse all pairwise argument similarities into one measure by taking the average following Equation (16).

$$\text{sim}_{\text{Aheads}}(p_1, p_2) = \frac{\sum_{\{a_1, a_2 | \text{label}(a_1) == \text{label}(a_2)\}} \text{sim}_{\text{WN}}(\text{head}(a_1), \text{head}(a_2))}{|\{a_1, a_2 | \text{label}(a_1) == \text{label}(a_2)\}|} \tag{16}$$

Relative Discourse Position. On the discourse level, we measure the distance of two predicate–argument structures with respect to their relative positions in the text. Given two predicates, we compute this measure based on the absolute difference between their relative positions. The relative position in discourse is computed as the *sentence_index* in which the predicate p_1 or p_2 occurs, divided by the total number of sentences in the affected document (d_1 or d_2, respectively). The measure, as defined in Equation (17), ranges from 1.0 (relative positions are exactly the same) to 0.0 (one predicate at the beginning, the other at the end of a text).

$$\text{sim}_{DPos}(p_1, p_2) = 1 - \left(\left| \frac{\text{sentence\_index}(p_1)}{\text{length}(d_1)} - \frac{\text{sentence\_index}(p_2)}{\text{length}(d_2)} \right| \right) \tag{17}$$

Context Similarity. We further consider occurrences of (shared) predicates in the immediate discourse context of two predicates. Our measure for this type of similarity is

computed as the relative number of overlapping predicate types within the preceding and succeeding n neighboring predicate instances as defined in Equation (18).

$$\begin{aligned} \text{sim}_{\text{DCon}}(p_1, p_2) &= \frac{\text{context}(p_1) \cap \text{context}(p_2)}{\text{context}(p_1) \cup \text{context}(p_2)}, \text{with} \\ \text{context}(p) &= \{p' | \text{index}(p') \in [\text{index}(p) - n : \text{index}(p) + n]\} \end{aligned} \tag{18}$$

We compute the index of a predicate as the number of preceding predicates within the same text, that is, starting with zero. Based on preliminary experiments on our development set, we empirically set n to 5.

4.2.3 Alignment via Clustering. The goal of this step of our induction framework is to identify pairs of PASs that describe the same event, state, or entity. As a means to achieve this goal, we represent pairs of comparable texts as graphs and aim to find those edges in a graph that represent connections between predicates that need to be aligned. Although our aim is to find edges between nodes, we note that the majority of predicates (nodes) in our data set are not aligned and hence a crucial prerequisite to generate precise alignments is to filter out those nodes that are unlikely to be good alignment candidates. To achieve the filtering and alignment goals at the same time, we rely on graph clustering techniques that have successfully been applied in the NLP literature (Su and Markert 2009; Cai and Strube 2010; Chen and Ji 2010, inter alia) and that can be used to partition a graph into singleton nodes and smaller subgraphs.

The clustering method applied in our model relies on so-called minimum cuts (henceforth also called **mincuts**) in order to partition a bipartite graph, representing pairs of texts, into clusters of alignable predicate–argument structures. A mincut operation divides a given graph into two disjoint subgraphs. Each cut is performed between some source node s and some target node t, such that (1) each of the two nodes will be in a different subgraph and (2) the sum of weights of all removed edges will be as small as possible. We implement basic graph operations using the freely available Java library JGraphT[7] and determine each mincut using the method described in Goldberg and Tarhan (1986).

Given a constructed input graph G, our algorithm recursively applies mincuts in three steps as follows:

1. Identify edge e with the lowest weight in the current (sub)graph G.
2. Perform a mincut such that the nodes connected by e will be in two different subgraphs G' and G''.
3. Recursively apply Steps 1 and 2 to subgraphs G' and G''.

As our goal is to induce clusters that correspond to pairs of corresponding structures, we apply Step 3 of the clustering approach outlined above only to subgraphs that contain more than two nodes. Algorithm 1 provides a pseudocode implementation of this procedure. An example of an input graph and the applied minimum cuts is illustrated in Figure 1. As shown in the illustration, we only use edges in our initial graph representation that represent alignment candidates with a similarity above a

7 http://jgrapht.org/.

Algorithm 1. Pseudo code of our clustering algorithm.

```
function CLUSTER(G)
    clusters ← ∅
    E ← GETEDGES(G)                                   ▷ Step 1
    e ← GETEDGEWITHLOWESTWEIGHT(E)
    s ← GETSOURCENODE(e)
    t ← GETTARGETNODE(e)
    G' ← MINCUT(G, s, t)                              ▷ Step 2
    C ← GETCONNECTEDCOMPONENTS(G')
    for all G_s ∈ C do                                ▷ Step 3
        if SIZE(G_s) <= 2 then
            clusters ← clusters ∪ G_s
        else
            clusters ← clusters ∪ CLUSTER(G_s)
        end if
    end for
    return clusters;
end function
```

threshold determined on the development part of our manually aligned data. Based on the initial representation, the first cut (Cut 1) is performed between the nodes connected by the edge with the lowest weight in the overall graph (13.0). This cut separates the nodes representing the "earnings_NNS" predicates from the rest of the graph. Similarly, Cut 2 separates another cluster of two nodes. Finally, Cuts 3 and 4 remove a single node from the only remaining cluster that had more than two nodes.

The main benefit of our method compared with off-the-shelf clustering techniques is that we can define the termination criterion in line with the goal of our task, namely,

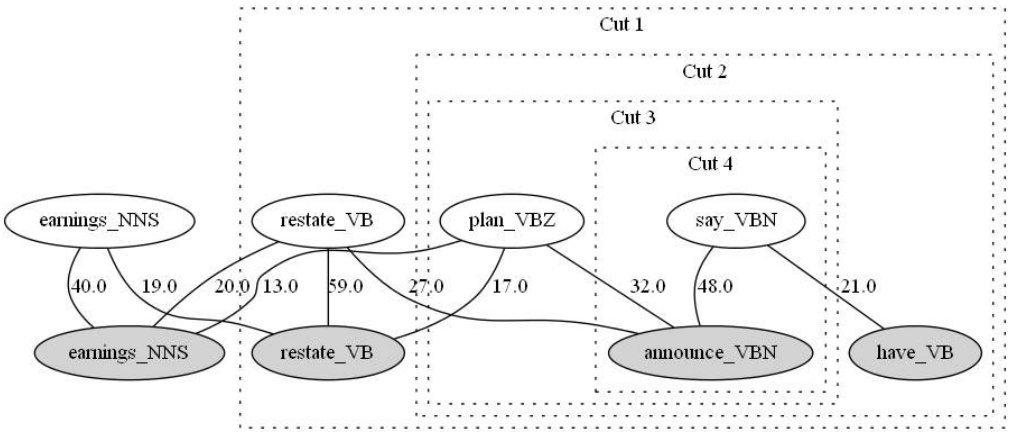

Figure 1
The predicates of two sentences (white: "The company has said it plans to restate its earnings for 2000 through 2002."; grey: "The company had announced in January that it would have to restate earnings (...)") from the Microsoft Research Paragraph Corpus are aligned by computing clusters with minimum cuts.

to align *pairs* of structures across comparable texts, instead of having to optimize additional parameters that require careful fine-tuning (such as the number of clusters or a clustering threshold). In the next section, we provide empirical evidence for the advantage of this approach.

4.3 Experiments

This section describes the evaluation of our graph-based clustering model on the task of aligning predicate–argument structures across comparable texts. We define two variants of our graph-based model: **Full** makes use of all similarity measures described in Section 4.2.2, and **EMNLP'12**, the model introduced in Roth and Frank (2012b), only uses the predicate-specific measures and the bag-of-words measure for argument similarity. Both models, **Full** and **EMNLP'12**, make use of the clustering algorithm introduced in Section 4.2.3.

For evaluation, we compare our model against two baselines and a model from the literature that has recently been proposed for this task (Wolfe et al. 2013). Similar to our approach, Wolfe et al. use various resources to calculate the similarity of two predicate–argument structures. Differences to our model lie in the kind of utilized resources, the use of additional data to learn feature weights, and the fact that each alignment decision is made using a binary classifier. We evaluate the predictions of each model on the manually annotated test set described in Section 4.1.2.[8]

As evaluation measures, we use precision, recall, and F_1-score. Following previous work aligning monolingual texts (Cohn, Callison-Burch, and Lapata 2008), we measure precision as the number of predicted alignments also annotated in the gold standard divided by the total number of predictions, and recall as the number of correctly predicted *sure* alignments divided by the total number of *sure* alignments in the gold standard. F_1-score is computed as as the harmonic mean between precision and recall.

4.3.1 Baselines. A simple baseline for predicate alignment is to simply cluster all predicates that have identical lemmata (henceforth called **LemmaId**). To assess the benefits of the clustering step, we propose a second baseline that uses the same similarity measures as our **Full** model but does not use the mincut clustering described in Section 4.2.3. Instead, it greedily merges as many 1-to-1 alignments as possible, starting with the highest similarity (**Greedy**). As a more sophisticated baseline, we make use of alignment tools commonly used in statistical machine translation. We train our own word alignment model using the state-of-the-art word alignment tool Berkeley Aligner (Liang, Taskar, and Klein 2006). As word alignment tools require pairs of sentences as input, we first extract paraphrases using a re-implementation of a previously proposed paraphrase detection system based on lemma and *n*-gram overlap (Wan et al. 2006). In the following section, we abbreviate the alignment based model as **WordAlign**.

4.3.2 Results. The results for the alignment task are presented in Table 2. From all approaches, **Greedy** and **WordAlign** yield the lowest performance. For **WordAlign**, we observe two main reasons. On the one hand, sentence paraphrase detection does not perform perfectly. Hence, the extracted sentence pairs do not always contain gold

8 We also performed an evaluation on sentence-level predicate alignment, but skipped the discussion here as this task is not relevant for our induction framework. As the additional discourse-level measures of the **Full** model are not needed for alignment within sentences, we refer the interested reader to the **EMNLP'12** model and its evaluation in Roth and Frank (2012b).

Table 2
Results for discourse-level alignment in terms of precision (P), recall (R), and F_1-score (all numbers in %); left: comparison of the **Full** model to baselines and previous work; right: impact of removing individual measures and using a tuned weighting scheme; results that significantly differ from **Full** are marked with asterisks (* $p < 0.05$; ** $p < 0.01$).

	P	R	F_1
	Baselines		
LemmaId	40.3**	60.3*	48.3**
Greedy	12.5**	27.6**	17.2**
WordAlign	19.7**	15.2**	17.2**
	Previous work		
Roth & Frank (2012b)	58.7**	46.6	52.0
Wolfe et al. (2013)	52.4**	**64.0****	57.6
	This article		
Full	**71.8**	48.9	**58.2**

	P	R	F_1
–sim$_{WN}$	78.8	44.6**	57.0
–sim$_{VN}$	78.7	44.6**	57.0
–sim$_{Dist}$	77.3	45.5**	57.3
–sim$_{ABoW}$	67.6*	49.3	57.0
–sim$_{Aheads}$	68.9	**52.9****	**59.8**
–sim$_{DPos}$	64.2	43.0**	51.6**
–sim$_{DCon}$	69.3*	52.2**	59.6
Full	71.8	48.9	58.2
+HighPrec	**86.2**	29.1**	43.5**

alignments. On the other hand, even sentence pairs that contain gold alignments can differ considerably in content and length, making them hard to align for statistical word alignment tools. The difficulty of this task can also be seen in the results for the **Greedy** model, which only achieves an F_1-score of 17.2%. In contrast, we observe that the majority of all sure alignments can be retrieved by applying the **LemmaId** model (60.3% recall).

The **Full** model achieves a recall of 48.9% but significantly outperforms all baselines ($p < 0.01$) in terms of precision (71.8%). This is an important factor for us as we plan to use the alignments in subsequent steps of our framework. With 58.2%, **Full** also achieves the best overall F_1-score. By comparing the results with those of the **EMNLP'12** model, we can see that the discourse-level similarity measures provide a significant improvement in terms of precision without a considerable loss in recall. This advantage in precision can also be seen in comparison to Wolfe et al. In contrast, their system outperforms our model with respect to recall. There are two main reasons for this: On the one hand, their model makes use of much larger resources to compute alignments, including a paraphrasing database that contains over 7 million rewriting rules; on the other hand, their model is supervised and makes use of additional data to learn weights for each of their features. In contrast, **Full** and **EMNLP'12** only make use of a small development data set to determine a threshold for graph construction. Though the difference is not significant, it is worth noting that our model outperforms that by Wolfe et al. by 0.6 percentage points in F_1-score, despite not making use of any additional training data.

Ablating Similarity Measures. All aforementioned results were conducted in experiments with a uniform weighting scheme of similarity measures as introduced in Section 4.2.2. Table 2 additionally shows the performance impact of individual similarity measures by removing them completely (i.e., setting their weight to 0.0). The numbers indicate that almost all measures contribute positively to the overall performance when using equal weights. Except for the argument head similarity, all ablation tests revealed significant

drops in performance, either with respect to precision or recall. This result highlights the importance of incorporating predicate-specific, argument-specific, and discourse-specific information regarding individual predications in this task.

Tuning Weights for High Precision. Subsequently, we tested various combinations of weights on our development set in order to estimate a better weighting scheme. This tuning procedure is implemented as a grid search technique, in which random weights between 0.0 and 1.0 are assigned to each measure. For graph construction, all weights are normalized again to sum to 1.0. We additionally try different thresholds for adding edges in the graph representation. To achieve high precision, we weight precision three times higher than recall while evaluating different parameters. In total, we tested 2,000 different parameter assignments on our development set. Following this process, we found the best result to be achieved with a threshold of 0.85 and the following weights:

- 0.11, 0.14, and 0.21 for sim_{WN}, sim_{VN}, and sim_{Dist}, respectively, (i.e., 46% of the total weight for predicate-specific measures)
- 0.21 and 0.05 for sim_{ABoW} and $\text{sim}_{\text{Aheads}}$, respectively, (i.e., 26% of the total weight for argument-specific measures)
- 0.21 and 0.07 for sim_{DPos} and sim_{DCon}, respectively (i.e., 28% of the total weight for discourse-specific measures)

The weighting scheme shows that information from all categories is considered. When applying the tuned model on our evaluation data set, we note that results in recall drop to 29.1% (-19.8 percentage points). Precision, on the other hand, increases to 86.2% ($+14.4$ percentage points).

4.4 Summary

In this section, we introduced the task of aligning predicate–argument structures across monolingual comparable texts. We designed annotation guidelines and created a data set of gold-standard alignments. Based on this data set, we developed and evaluated a novel clustering-based alignment model that uses a combination of various similarity measures and a graph-based clustering algorithm that we specifically designed for this task. In an intrinsic evaluation, we showed that our novel model outperforms a range of baselines as well as previous approaches to this particular task. As an additional contribution, we defined a tuning routine that can be utilized to train a high precision model for the discourse-level alignment task. Our results show that, by using this tuning step, corresponding structures in our evaluation set can be identified with a precision of 86.2%. This intermediate result is essential for the success of our overall framework. In the next section, we present Step 2 of our implicit argument induction technique, in which we examine pairs of automatically aligned predicate–argument structures as a means to identify and link implicit arguments.

5. Induction Framework Step 2: Extracting Implicit Arguments

In the second step of our induction framework, we rely on alignments between PAS to detect implicit arguments. That is, we aim to identify argument instances that are present, or *explicit*, in one PAS but absent, or *implicit*, from the aligned PAS. Based on the identified instances, our model tries to find antecedents of implicit arguments

within their respective discourse contexts. To perform this task automatically, we rely on several preprocessing tools, which we apply on the full corpus of over 160,000 document pairs introduced in Roth and Frank (2012a).

We describe the applied preparatory steps and the used preprocessing tools in Section 5.1. Using automatically computed annotations as input, we describe a heuristic method to detect implicit arguments and their discourse antecedents in Section 5.2.

5.1 Data Preparation

As a basis for the actual induction, we rely on several preparatory steps that identify information two documents have in common (cf. Figure 2). In particular, we compute and align PAS using the graph-based model described in Section 4, and determine co-referring entities across pairs of texts using coreference resolution techniques on concatenated document pairs (Lee et al. 2012). In theory, arguments implicit in one structure can straightforwardly be induced based on this information by looking for co-referring mentions of the argument explicit in the aligned structure. In practice, we make use of additional checks and filters to ensure that only reliable information is being used. We describe the preprocessing steps in the following paragraphs and provide additional details on our implementation of the induction procedure in Section 5.2.

Single Document Preprocessing. We apply several preprocessing steps to each document in our data set. First, we use the Stanford CoreNLP package (Manning et al. 2014) for tokenization and sentence splitting. We then apply the MATE tools (Björkelund et al. 2010; Bohnet 2010), including the integrated PropBank/NomBank-based semantic parser, to determine local PAS. Finally, we resolve pronouns that occur in a PAS using the coreference resolution system by Martschat et al. (2012), which placed second for English in the CoNLL-2012 Shared Task (Pradhan et al. 2012).

High Precision Alignments. Once all single documents are preprocessed, we align PAS across pairs of comparable texts. We want to induce reliable instances of implicit arguments based on aligned PASs pairs and hence apply our graph-based clustering technique using the high-precision tuning step described in Section 4.3. We run

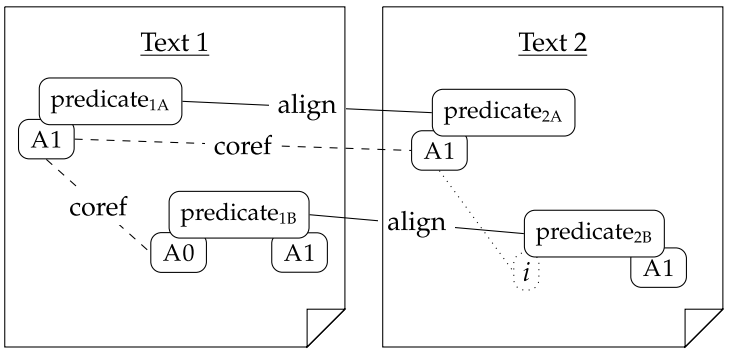

Figure 2
Illustration of the induction approach: Texts consist of PAS (represented by overlapping rounded rectangles); we exploit alignments between corresponding predicates across texts (solid lines) and co-referring entity mentions (dashed lines) to infer implicit arguments (marked by '*i*') and link antecedents (dotted line).

Table 3
Properties of the high precision alignment data set.

	high-precision alignments	
number of alignments	283,588	
same POS	278,970	(98.4%)
noun–noun	89,696	(31.6%)
verb–verb	189,274	(66.8%)
mixed POS	4,618	(1.6%)
same lemma	273,924	(96.6%)
different lemma	9,664	(3.4%)
same number of arguments	239,563	(84.5%)
unequal number of arguments	44,025	(15.5%)

the high-precision model on all pairs of texts in the corpus. As a result, we extract a total number of 283,588 aligned pairs of PASs. An overview of properties of this data set is given in Table 3. We observe that most alignments involve the same part-of-speech (98.4%) and the same predicate lemma (96.6%). This fact simplifies the task of inducing implicit arguments as it implies that in most cases the PropBank argument labels in a pair of aligned structures correspond to each other and do not have to be mapped via a higher-level set of roles.

Cross-Document Coreference. For each argument that is explicit in one PAS but implicit in an aligned PAS, we want to determine a suitable antecedent within the discourse context of the implicit instance. We solve this task by viewing the aligned explicit argument as an entity reference and identify co-referring mentions in both texts by applying coreference resolution techniques across pairs of documents. In practice, we follow the methodology of Lee et al. (2012), who propose applying standard coreference methods on pairs of texts by simply concatenating two documents and providing them as a single input document. As merging two different discourses can lead to problems for a coreference system that is based on features and feature weights trained on single documents, we follow Lee et al. and apply a rule-based system. Like them, we use the Stanford Coreference system (Lee et al. 2013), which applies a sequence of coreference "sieves" to the input, ordered according to their precision. To obtain a highly accurate and reliable output, we consider only the most precise resolution sieves:

- "String Match"
- "Relaxed String Match"
- "Precise Constructs"
- "Strict Head Match A," "Strict Head Match B," "Strict Head Match C"
- "Proper Head Noun Match"

Note that none of these sieves involve pronoun resolution. Instead, we decided to use the resolved pronouns from the single-document coreference step. This decision

is based on the fact that the system by Martschat et al. (2012) outperforms the Stanford system with all sieves on the CoNLL'11 test set by an average F_1-score of 3.0 absolute points. The high-precision sieves are, however, better suited for the cross-document task as we plan to rely on the resulting coreference chains for identifying potential antecedents of implicit arguments. That is, we prefer fewer but more reliable chains in order to minimize the impact of possible preprocessing errors.

5.2 Automatic Identification and Linking

Given a pair of aligned predicates from two comparable texts, we examine the output of the semantic parser (cf. Section 5.1) to identify arguments in each PAS. We compare the set of labels assigned to the arguments in each structure to determine whether one PAS contains an argument (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 implicit arguments within discourse, we require candidate antecedents to be mentions that occur in the same document. We impose a number of restrictions on the resulting pairs of implicit arguments and antecedents to reduce the impact of different types of preprocessing errors:

Mislabeled Arguments. In some cases, the parser annotated the same argument in two texts using different labels. To ensure that mislabeled arguments are not recognized as implicit, we require that pairs of aligned PAS contain a different number of arguments.

Missed Arguments. Depending on sentence structure, the semantic parser is sometimes unable to determine all local arguments. This often leads to the identification of erroneous implicit arguments. To intercept some of these cases, we require that all antecedents from the cross-document coreference chain must be outside of the sentence that contains the affected PAS.

5.3 Resulting Data Set

We apply the outlined identification and linking approach to all text pairs in our corpus of comparable texts. As a result, we induce a total of 698 implicit argument and antecedent pairs. A summary of properties of the obtained pairs can be found in Table 4. The full data set involves predicates of 294 different lemmata. Each pair was found in a separate document. Note that 698 implicit arguments from 283,588 pairs of PAS seem to represent a fairly low recall. The reason for this is that most PAS pairs in the high-precision data set consist of identically labeled argument sets (84.5%) and in most of the remaining cases an antecedent in discourse cannot be found using the high-precision coreference sieves. This does not mean that implicit arguments are rare in general. As discussed in Section 4.1, 38.9% of all manually aligned PAS pairs involve a different number of arguments.

We manually evaluated a subset of 90 induced implicit arguments and found 80 discourse antecedents to be correct (89%). Examples are displayed in Table 5. A closer analysis revealed that some incorrectly linked instances still result from preprocessing

Table 4
Properties of automatically induced implicit arguments and antecedents.

	implicit arguments and discourse antecedents	
number of induced instances	698	
predicate counts		
nominal predicates	285	(40.8%)
verbal predicates	413	(59.2%)
word/lemma types	535/294	
antecedent distance to predicate		
previous sentence	240	(34%)
within the previous 5 sentences	415	(59%)
within the previous 10 sentences	442	(63%)
follow-up context	260	(37%)
label of induced argument		
proto-agent (A0)	423	(60.6%)
proto-patient (A1)	107	(15.3%)
other (A2–A5)	168	(24.1%)

errors. In particular, combinations of errors can lead to incorrectly identified instances as showcased in Example 19:

Example 19
"The Guatemalan Congress on Thursday ratified 126-12 [a Central America-US]$_{A0}$ [free trade]$_{A1}$ agreement, lawmakers said."
Induced missing argument and discourse antecedent: [goods]$_{A2/co\text{-}agent}$

Instead of recognizing *Central America* and *US* as two separate arguments (agent and co-agent) of the predicate AGREE, the semantic parser labels both entities as one argument (A0, agent); our system hence tries to determine a discourse antecedent for an argument that is predicted to be missing despite being actually realized (A2, co-agent). In the aligned PAS, the co-agent is realized as a prepositional phrase: "[with the United States]$_{A2}$". The cross-document coreference tool incorrectly predicts *the United States* in this phrase to be coreferent with the phrase *U.S. goods and services*; our system hence detects *goods* as the antecedent for the erroneously predicted implicit argument.

Further error sources are incorrectly extracted document pairs and alignments between PAS that do not correspond to each other. Example 20 shows two PAS from a pair of texts that describe different sets of changes in economic activity.

Example 20
"[Production]$_{A1}$ rose [3.9 percent from October in the capital good sector]$_{A2}$"
"[Industrial production excluding energy, food and construction ...]$_{A1}$ rose [2.6 percent]$_{A2}$ [in November]$_{TMP}$ from the previous month"

Example 21 shows a pair of aligned structures in a pair of comparable texts that both report on two planned trips by President Obama. Here, the alignment model erroneously aligned a reference to one trip with a reference to both trips.

Table 5
Three positive examples of automatically induced implicit arguments (∅) and the cross-document coreference chains that include discourse antecedents (**i**); the right-hand side shows the aligned PAS that were used to identify a suitable antecedent for the implicit argument in the text on the left-hand side.

[T-Online$_{\mathbf{i}}$], the leading Internet services provider in Europe and a unit of Deutsche Telekom, said Thursday its net loss more than doubled last year owing to its foreign activities and goodwill writedowns. (. . .) The [∅$_{\mathbf{i}}$]$_{A0}$ [operating]$_{A3}$ **loss**, as measured by earnings before interest, tax, depreciation, and amortization, widened to 189 million euros last year from 121.6 million euros a year earlier.	[T-Online's$_{\mathbf{i}}$]$_{A0}$ [operating]$_{A3}$ **loss** — earnings before financial items such as interest, taxes, depreciation, and amortization — also widened, to 189 million euros (dlrs 167 million) in 2001 from 122 million (dlrs 108 million).
[Mozambique$_{\mathbf{i}}$] police have arrested four foreigners in connection with an alleged plot to sabotage the African country's largest hydroelectric dam, officials said Wednesday. (. . .) It was handed over to Mozambican control last year, 33 years after [∅$_{\mathbf{i}}$]$_{A1}$ **independence** [in 1975]$_{TMP}$.	Its power lines and other infrastructure sustained severe damage during the 16-year civil war that followed [Mozambique's$_{\mathbf{i}}$]$_{A1}$ **independence** [in 1975]$_{TMP}$.
The accident occurred just after midnight on Sunday in Shanxi province but [local officials]$_{A0}$ failed to immediately **report** [the accident]$_{A1}$ [∅$_{\mathbf{i}}$]$_{A2}$, the State Administration for Work Safety said on its website. [The government$_{\mathbf{i}}$] says 4,750 people died in coal mine accidents last year, an average of 13 a day. It is common for mine owners to delay reporting accidents or to not report them at all.	The explosion happened in a mine in the suburbs of Jincheng City on Sunday in Shanxi province, but [the coal mine owner]$_{A0}$ did [not]$_{NEG}$ [immediately$_{TMP}$] **report** [it]$_{A1}$ [to the government$_{\mathbf{i}}$]$_{A2}$, Xinhua News Agency said.

Example 21
"The postponement ... marks the second time [the president's]$_{A0}$ trip [to Indonesia]$_{A1}$ has been postponed"
"Obama was to depart on a [weeklong]$_{TMP}$ trip [to both countries . . .]$_{A1}$ on June 13"

5.4 Summary

In this section, we introduced a computational implementation of Step 2 of our rule-based induction method for identifying implicit arguments and their discourse antecedents. Our approach depends on automatic annotations from semantic role labeling, PAS alignment, and coreference resolution. We implement two particular types of measures to minimize the impact of preprocessing errors: (1) we avoid imprecise input by applying high-precision tools instead of methods that are tuned for balanced precision and recall; and (2) we circumvent some common preprocessing errors by formulating

constraints on resulting instances of implicit arguments and discourse antecedents. An intrinsic analysis reveals that we cannot eliminate all error sources this way, although we found the induced data set to be of high precision.

The next sections present extrinsic evaluations in which we demonstrate the utility of the induced instances empirically. As discussed in Section 1, we assume that a proper treatment of implicit arguments can improve the coverage of semantic role labeling systems and entity-based models of coherence. We show how our data set can be utilized in both of these areas: In Section 6, we use our data set of implicit arguments to enhance the training process for models of implicit argument linking; in Sections 7, we demonstrate that implicit arguments can affect the perceived coherence of a text and that instances of aligned (explicit and implicit) arguments in our data set can be used to successfully predict this impact adequately.

6. Task Setting 1: Linking Implicit Arguments in Discourse

Our first extrinsic evaluation assesses the utility of automatically induced pairs of implicit arguments and antecedents for the task of implicit argument linking. For this scenario, 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 task*). For direct comparison with previous results and heuristic acquisition techniques, we apply the implicit argument identification and linking model by Silberer and Frank (2012, henceforth S&F) for training and testing. We briefly describe the SemEval task data and the model by S & F in the next sections.

6.1 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"), gold labels are provided for local arguments and participating systems have to perform the following three sub-tasks: firstly, non-instantiated roles have to be identified; secondly, they have to be classified as being "accessible to the speaker and hearer" (definite null instantiation, DNI) or as being "only existentially bound within discourse" (indefinite null instantiations, INI); and finally, all resolvable null instantiations have to be linked to discourse antecedents.

The task organizers provide two versions of their data set: one based on FrameNet annotations and one based on PropBank/NomBank annotations. We found, however, that the latter 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 automatically induced data set, which we want to apply as additional training data, is labeled with a PropBank/NomBank-style parser. Consequently, we need to map our annotations to FrameNet in order to make use of them in this task. The organizers of the SemEval task provide a manual mapping dictionary for predicates in the annotated data set. We use this manual mapping and additionally use SemLink (Palmer 2009) for mapping predicates and arguments not covered by the dictionary.

6.2 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 subsection, this task consists of three steps: In Step (1), implicit

arguments are identified as unfilled and non-redundant FrameNet core roles; in Step (2), an SVM classifier is used to predict whether implicit arguments are definite based on a small number 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 to find appropriate antecedents for (predicted) definite arguments. S&F report that their best results were obtained considering the following target antecedents: all entity mentions that are syntactic constituents from the present and the past two sentences and all entity mentions 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 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. To find the best set of features, we run the feature selection on a combination of the SemEval training data and our full additional data set.[9] The only features that were selected in this process concern the "prominence" of the candidate antecedent, its semantic agreement with the selectional preferences of the predicate, the part-of-speech-tags used in each reference to the candidate entity, and the semantic types of all roles that the entity fills according to local role annotations. These features are a subset of the best features described in Silberer and Frank (2012).

6.3 Results

Evaluation measures. For direct comparison in the full task, both with S&F's model and other models, we adopt the precision, recall, and F_1 measures as defined in Ruppenhofer et al. (2010).

Baselines. We compare our results with those previously reported on the SemEval task (see Table 6 for a summary): The best performing system in the actual task in 2010 was developed by Chen et al. (2010) and is an adaptation of the semantic role labeling system SEMAFOR (Das et al. 2010). In 2011, Tonelli and Delmonte (2011) presented a revised version of their SemEval system (Tonelli and Delmonte 2010), which outperforms SEMAFOR in terms of recall (6%) and F_1-score (8%). The best results in terms of recall and F_1-score 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 achieves a lower 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.

Our Results. Comparison of our results with those 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 of one percentage point in recall is marginal given the size of the test set (only 259 implicit arguments have an annotated antecedent). Our

9 Note that this feature selection procedure is the same as applied by Silberer and Frank. Hence, the evaluated models are directly comparable and all differences in results can directly be traced back to the use of additional data.

Table 6
Results for identifying and linking implicit arguments in the SemEval test set.

	Precision	Recall	F_1-score
Chen et al. (2010)[10]	**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, Ruppenhofer, and Sporleder (2013)[11]	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 article	0.21	0.08	0.12

result in precision is the second highest score reported on this task. Interestingly, the improvements are higher than those achieved in the original study by Silberer and Frank (2012), even though their best additional training set is three times bigger than ours and contains manual semantic annotations.

We conjecture that their low gain in precision could be a side effect triggered by two different factors. Firstly, the heuristically created training instances, induced by treating anaphoric pronouns as implicit argument instances, might not reflect the same properties as actual implicit arguments. For example, pronouns can occur in syntactic constructions in which an actual argument cannot be omitted in practice, leading an incorrect overgeneralization. An additional negative factor might be that their model relies on coreference chains, which are automatically generated for the test set and hence rather noisy. In contrast, our automatically induced data does not contain manual annotations of semantic roles and coreference chains, hence we do not rely on gold information during training and testing. The results show that, despite this limitation, our new model outperforms previous models trained using the same system, indicating the utility and high reliability of our automatically induced data.

Impact of training data size. To assess the impact of training data size, we perform an additional experiment with subsets of automatically induced implicit arguments. Specifically, we train different classifiers using the full SemEval training data and varying amounts of our automatically induced training data (random samples of 1%, 10%, and 25%). The model uses the best feature set determined on the combination of SemEval and our full additional data set in all settings. For each setting, we report average results obtained over four runs. The outcomes of this experiment are summarized in Table 7. The numbers reveal that using only 1% of the additional data for training already leads to a classification performance of 0.13 in F_1-score. The large improvement over the S&F model without additional data, which achieves an F_1-score of 0.07, can be explained by the fact that the features selected by our model generalize better than those selected on the SemEval training data only. The improvements are highest when using between 10% and 25% of the additional data, indicating that the use of additional induced instances indeed increases performance but that utilizing more out-of-domain than in-domain data for training seems to be harmful.

10 Results as reported by Tonelli and Delmonte (2011).
11 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.

Table 7
Results for identifying and linking implicit arguments using features selected on our full data set and different combinations of task-specific and automatically induced data for training.

Training data	Precision	Recall	F_1-score
SemEval + induced instances (1%)	0.22	0.09	0.13
SemEval + induced instances (10%)	**0.24**	**0.10**	**0.14**
SemEval + induced instances (25%)	**0.24**	**0.10**	**0.14**
SemEval + induced instances (100%)	0.21	0.08	0.12

6.4 Summary

In this section, we presented a NLP application of the automatically induced data set of implicit arguments that we introduced in Section 5. We found that automatically induced implicit arguments can successfully be used as training data to improve a system for linking implicit arguments in discourse. Although the presented model cannot compete with state-of-the-art systems, the addition of our data led to an enhanced performance compared with the same system with different and without additional training data. Compared with the model without additional training data, our induced data set increased results in terms of precision and F_1-score by 15 and 5 percentage points, respectively.

7. Task Setting 2: Modeling Local Coherence

In our second experiment, we examine whether argument realization decisions affect the perceived coherence of a text and investigate how and which factors related to their impact can be used to model realization decisions computationally. We approach this question in the following way: We exploit PAS alignments across comparable documents to identify contexts with implicit and explicit arguments; we then make use of these automatically induced contexts in order to train a discourse coherence model that is able to predict whether—in a given context—an argument should be realized or remain implicit.

Induction of such a model and its evaluation will be approached as follows: First, we assemble a data set of document pairs that differ only with respect to a single realization decision; given each pair in this data set, we ask human annotators to indicate their preference for the implicit or explicit argument instance in the prespecified context (Section 7.1); secondly, we attempt to emulate the decision process computationally using a discriminative model based on discourse and entity-specific features (Section 7.2). To assess the performance of the new model, we train it on automatically induced training data and evaluate it, in comparison with previous models of local coherence, against human annotations (Section 7.3).

7.1 Data Compilation

We use the data set of automatically induced implicit arguments (henceforth *source data*), described in Section 5, 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 (22) illustrates an original and modified sentence.

Example 22

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

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

Note that adding or removing arguments at random can lead to structures that are semantically implausible. Hence, we consider two restrictions. First, we ensure that the remaining context is still understandable after an argument is removed by only considering texts in which follow-up references can still be resolved based on earlier antecedents. Second, we only modify arguments of PAS that actually occur and are aligned across two texts. Given a pair of PAS that differ with respect to an argument realization, we create modifications by replacing the specific implicit or explicit argument in one text with the corresponding argument in the paired text. Examples (22) and (23) show two such comparable sentences. The original PAS in Example (22a) contains an explicit argument that is implicit in the aligned PAS and hence removed in the modified version. Similarly, the original text in (23a) involves an implicit argument that is made explicit in the modified version (23a').

Example 23

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

(a') [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 pairs of PASs 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. In other words, 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, Grosz, and Gilliom 1993).

Annotation. We set up a Web experiment using the evaluation toolkit by Belz and Kow (2011) to collect ratings of local coherence for implicit and explicit arguments. For this experiment, we compiled a data set of 150 document pairs. Each text in such a pair consists of the same content, with the only difference being one argument realization.

We presented all 150 pairs to two annotators[12] 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 3. 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 (explicit/implicit). Instead of

12 Both annotators are native speakers of English and undergraduate students in literature and linguistics.

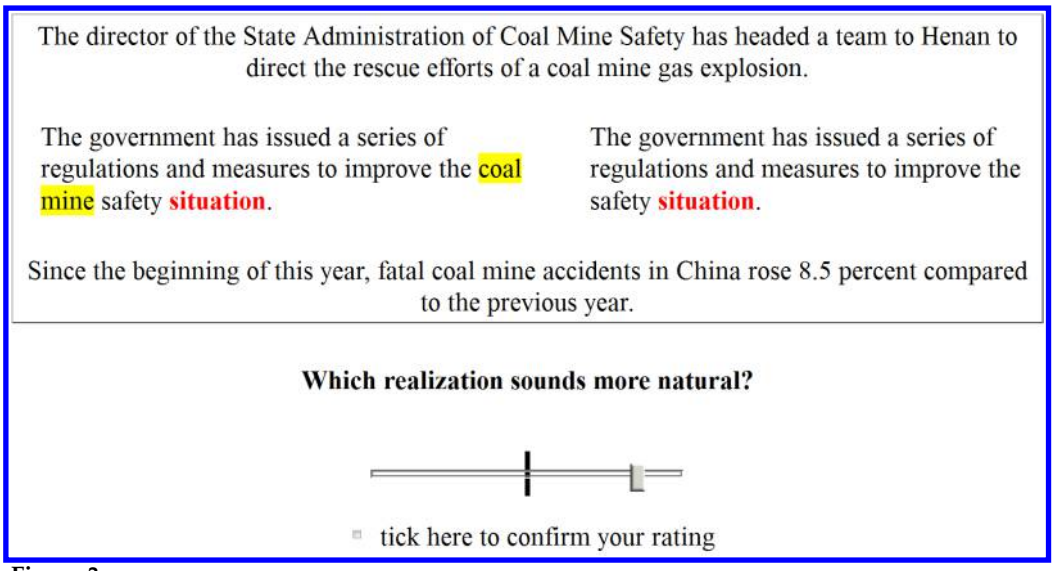

Figure 3
Texts as displayed to the annotators.

providing a definition of local coherence ourselves, we asked annotators to rate how "natural" a realization reads given the discourse context. This procedure is in line with previous work by Pitler and Nenkova (2008), who "view text readability and text coherence as equivalent properties," given that the annotators are "competent language users."

We found that annotators made use of the full rating scale, with the extremes indicating either a strong preference for the text on the left-hand side or the right-hand side, respectively. However, most ratings were 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 annotators affirmed that some cases do not read naturally when a specific argument is omitted or redundantly realized at a given position in discourse. For example, the text fragment in Example (24) shows two sentences in which an argument has been realized twice, leading to a perceived redundancy in the second sentence (A4, `destination`); conversely, Example (25) showcases an excerpt in which a non-redundant argument (A2, `co-signer`) has been omitted.

Example 24
The remaining contraband was picked up at Le Havre.
The containers had arrived [in Le Havre] from China.

Example 25
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 [∅].

We computed correlation between ratings of both annotators using Spearman's ρ (Spearman 1904) and found a low but significant correlation ($\rho = 0.22$, $p < 0.01$). To

Table 8
Statistics on the collected data and final test set.

Total number of instances	150		**Test instances**	70	
Preference			**Argument label**		
Explicit	47	(31%)	A0	24	(34%)
Implicit	23	(15%)	A1	9	(13%)
No preference	42	(28%)	A2	29	(41%)
Disagreement	38	(25%)	A3	3	(4%)
			A4	3	(4%)

construct a test set with gold annotations, we mapped continuous ratings to discrete labels (implicit, explicit, neutral) and selected a subset of 70 instances for which clear[13] preferences were observed. Table 8 provides some statistics on the collected data. Even though correlation with respect to continuous ratings is relatively low, we find that both annotators have the same general preference in most cases, with a raw agreement of 75% overall. Our test set contains 47 (31%) cases of explicit arguments and 23 (15%) cases of implicit arguments. Those cases in which annotators disagreed (25%) or had no preference (28%) were discarded.

7.2 Coherence Model

We model the decision process that underlies the (non-)realization of arguments using an SVM-based model and a range of discourse features. The features are based on the following three factors: the affected predicate–argument structure (**Parg**), the (automatic) coreference chain of the affected entity (**Coref**), and the discourse context (**Disc**).

Parg. The first group of features is concerned with the characteristics of the affected PAS: This includes the absolute and relative number of explicitly realized arguments in the structure, the number of modifiers in it, and the total length of the structure as well as of the complete sentence (in words).

Coref. The coreference-specific features include transition patterns as inspired by the entity grid model (cf. Section 3), the absolute number of previous and follow-up mentions of the (non-)realized argument, their POS tags, and the distance between the current PAS to the closest previous and follow-up mention (in number of words and sentences). In contrast to previous work on the entity grid model, we do not type transition features 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 training data (see the following).

Disc. On the discourse level, we define a small set of additional features that include the total number of coreference chains in the text, the occurrence of pronouns in the current sentence, lexical repetitions in the previous and follow-up sentence, the current position in discourse (begin, middle, end), and a feature indicating whether the affected argument occurred in the first sentence.

13 Absolute continuous rating of >1 standard deviation from zero.

Most of these features overlap with those successfully applied in previous work on modeling coherence. For example, the transition patterns are inspired by the entity grid model. In addition to entity-grid like features, Pitler and Nenkova (2008) also use text length, 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 PAS. In the evaluation (cf. Section 7.3), we report results for two models: A **simplified model** that only makes use of entity-grid like features and a **full model** that uses all features described here. To learn feature weights, we make use of the training data described in the following section.

7.3 Experiments

The goal of our model is to correctly predict the realization type (implicit or explicit) of an argument that maximizes the perceived coherence of the document. As a proxy for coherence, we use the naturalness ratings given by our annotators. We evaluate classification performance on the 70 data points in our annotated test set for which clear preferences have been established. We report results in terms of precision, recall, and F_1-score per class as well as micro- and macro-averaged F_1-score across classes. We compute precision as the fraction of correct classifier decisions divided by the total number of classifications made for a specific class label; recall as the fraction of correct classifier decisions divided by the total number of test items with the specific label; and F_1 as the harmonic mean between precision and recall.

7.3.1 Baselines. For comparison, we apply a couple of coherence models proposed in previous work: 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.[14] Note that the toolkit only returns one coherence score for each document. To use the model scores for predicting argument realization, 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 count the test item neither as correctly nor as incorrectly classified.

7.3.2 Our Models. Like the applied baseline models, our models do not make use of any manually labeled data for training. Instead, we utilize the automatically identified instances of explicit and implicit arguments from pairs of comparable texts, which we described in Section 5. To train our own model, we prepare this data set as follows: Firstly, we remove all data points that were selected for the test set; secondly, we split all pairs of texts into two groups—texts that contain a PAS in which an implicit argument has been identified (IA), and their comparable counterparts, which contain the aligned PAS with an explicit argument (EA). All texts are labeled 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

14 `http://www.ling.ohio-state.edu/%7Emelsner/`.

Table 9
Results for correctly predicting argument realization. Significant differences from our (full) model in terms of micro-averaged F_1-score are marked with asterisks (* $p < 0.01$).

Model	implicit argument P	R	F_1	explicit argument P	R	F_1	averaged F_1-scores macro	micro
Entity grid models	–	–	–	–	–	–	–	–
Baseline entity grid	0.30	0.78	0.43	0.44	0.09	0.14	0.29	0.31*
Extended entity grid	0.30	0.78	0.43	0.50	0.11	0.18	0.30	0.33*
Topical entity grid	0.31	0.87	0.45	0.40	0.04	0.08	0.27	0.31*
Other models	–	–	–	–	–	–	–	–
Pronouns	0.30	0.48	0.37	0.58	0.23	0.33	0.35	0.35*
Discourse-newness	0.35	**0.96**	0.52	**0.88**	0.15	0.25	0.39	0.41*
This article	–	–	–	–	–	–	–	–
Simplified model	0.70	0.30	0.42	0.73	**0.94**	0.82	0.66	0.73
Our (full) model	**0.67**	0.43	**0.53**	0.76	0.89	**0.82**	**0.69**	**0.74**

the realization decision had not been performed and can thus extract unbiased feature values for the affected entity and argument position. Given each feature representation, we train a classifier using the default parameters of the LIBSVM package (Chang and Lin 2011).[15]

We apply our own model on each data point in the small annotated test set, where we always treat the affected argument, regardless of its actual annotation, as implicit to extract unbiased feature values for classification. Based on the features described in Section 7.2 and trained on the automatically constructed PAS alignments, our model predicts the realization type of each argument in the given context. We note that our model has an advantage here because it is specifically designed for this task and trained on corresponding data. All models compute local coherence ratings based on entity occurrences, however, and should thus be able to predict which realization type coheres best with the given discourse context. That is, because the input document pairs are identical except for the affected argument position, the coherence scores assigned by each model to pairs of text only differ with respect to the affected entity realization.

7.3.3 Results. The results are summarized in Table 9. We begin our analysis with a discussion on results in terms of micro-averaged F_1-scores. As observable from the last column in Table 9, each of the baseline models achieves an F_1-score between 31% and 41%, whereas our (full) model achieves 74%. As our data set is imbalanced and biased towards explicit realizations, we also provide macro-averaged F_1-scores that show average performance over both classes without taking into account their respective sizes. Here, we find that baseline results lie between 27% and 39%, whereas our (full) model achieves 69%. Before discussing the results of our model in more detail, we investigate baselines performances.

We observe that the original entity grid model exhibits a preference for the implicit realization type: It predicts this class in 61 (87%) cases, resulting in only 9% of all explicit arguments being correctly classified. Overall, the entity grid achieves a

15 The default settings in our LIBSVM version are: equal costs of both classes and use of a sigmoid kernel.

(micro-averaged) F_1-score of 31%. Taking a closer look at the features of the model reveals that this is an expected outcome: In its original setting, the entity grid learns realization patterns in the form of sentence-to-sentence transitions. As discussed in the paper that introduced the entity grid model (Barzilay and Lapata 2008), most entities are, however, only mentioned a few times in a text. Consequently, non-realizations constitute the "most frequent" 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, suffer less from such an effect and achieve better results, with F_1-scores up to 35% and 41%, respectively. The extended entity grid model also alleviates the bias towards non-realizations, resulting in a slightly improved F_1-score of 33%. To abstract away from this issue, we train a simplified version of our own model that only uses features that involve entity transition patterns. The main difference between this simplified model and the original entity grid model lies in the different use of training data: whereas 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 confirm that this extra information has a significant impact ($p < 0.01$, using a randomization test; Yeh 2000) on test set performance, and raises the ratio of correctly classified explicit arguments to 73%. Yet, the simplified model only provides a correct label for 30% of instances in the test data that are marked as implicit arguments, with a class-specific F_1-score of only 42%. As demonstrated by the performance of our full model, a combination of all features is needed to achieve the best overall results of 69% and 74% in macro and micro-averaged F_1-scores, respectively. Applied to the two classes separately, our model achieves an F_1-score of 53% on arguments that are annotated as implicit and 82% on explicit arguments. Both of these scores are the best across all tested models.

To determine the impact of the three different feature groups, we derive the weight of each feature from the model learned by LIBSVM. Table 10 gives an overview over the ten highest weights for implicit and explicit realization classification. We use the following terminology in the feature description: "the entity" refers to the entity that is referred to by the to-be-classified argument, "next/previous mention" denotes a co-referring mention to the same entity, "the PAS" refers to the predicate–argument structure that contains the affected argument (implicitly), and "the sentence" refers to the sentence in which this PAS is realized. All "distances" refer to the number of tokens that appear between the predicate that heads the PAS and the previous or next mention of the entity. As can be seen at the top and bottom ends of Table 10, the strongest feature for classifying an argument as implicit is whether the entity is also realized in the preceding or following two sentences. The strongest feature for classifying an argument as explicit is whether the next mention is a pronoun. The trained weights indicate that the model is learning some interesting patterns that reflect rules such as "avoid close repetitions," "keep sentences short," and "pronouns and proper names can more often be dropped than definite noun phrases."

7.4 Summary

In this chapter, we presented a computational linguistic application of the automatically induced data set of implicit arguments that we introduced in Section 5. This data set has been induced from pairs of comparable text and is a unique resource in that it contains automatic annotations of implicit arguments, aligned explicit arguments, and discourse antecedents. In our experiments on perceived coherence, we found that the

Table 10
Weights assigned to each feature in our model; list includes the top 10 features for implicit (positive weight) and explicit arguments (negative).

Weight	Group	Feature description
+55.38	Coref	The entity is mentioned within two sentences
+25.37	Coref	The entity has previously been mentioned as a proper noun
+19.14	Coref	The entity has previously been mentioned as a pronoun
+14.75	Parg	The PAS consists of at least 2 words
+12.82	Parg	The sentence contains at least 20 words
+12.65	Parg	The sentence contains at least 40 words
+12.12	Parg	The PAS consists of at least 3 words
+11.32	Coref	The entity is mentioned in the next but not in the previous sentence
+11.23	Coref	The entity is mentioned within the previous or next 10 tokens
+10.79	Coref	The entity is mentioned within the previous two sentences
−4.72	Parg	The absolute number of arguments and modifiers in the PAS
−5.80	Coref	The entity is mentioned two sentences ago but not in the previous
−6.38	Parg	The previous entity mention was a definite noun phrase
−6.94	Disc	The PAS occurs in the first sentence of the discourse
−7.11	Parg	The absolute number of arguments in the affected PAS
−7.22	Coref	The entity is mentioned in the next sentence but not in the previous
−8.79	Coref	The entity is mentioned within the previous or next three sentences
−9.42	Coref	The entity is mentioned within the previous three sentences
−10.38	Coref	The entity is mentioned in the previous sentence
−32.70	Coref	The next mention is a pronoun

use of implicit vs. explicit arguments, although often being a subtle difference, can have a clear impact on readability ratings by human annotators. We showed that our novel coherence model, which is solely trained on automatically induced data, is able to predict this difference in newswire articles with an F_1-score of up to 74%.

8. Discussion and Conclusions

In this article, we introduced a framework for inducing instances of implicit arguments and their discourse antecedents from pairs of comparable texts, and showcased applications of these instances in natural language processing. As described in Section 2, our framework consists of two steps: aligning predicate–argument structures across comparable texts and identifying implicit arguments and antecedents. In the following paragraphs, we summarize our framework and highlight specific contributions.

Predicate–Argument Structure Alignments. With the goal of inducing instances of implicit arguments, we proposed a novel task that aims to align pairs of PASs across pairs of comparable texts. In Section 4, we introduced a manually annotated data set for the development and evaluation of models for this particular task. We found that pairs of PASs can be aligned across documents with good inter-annotator agreement given appropriate annotation guidelines. Based on the development part of our corpus, we designed and fine-tuned a novel graph-based clustering model. To apply this model, we represent PAS in pairs of documents as bipartite graphs and recursively divide this graph into subgraphs. All clustering decisions by the model are based on pairwise similarities between PASs, combining information on predicates, associated arguments,

and their respective discourse contexts. We empirically evaluated our model against various baselines and a competitive model that has recently been proposed in the literature. The results of our evaluation show that our model outperforms all other current models on the PAS alignment task by a margin of at least 0.6 percentage points in F_1-score, despite only a single threshold parameter being adjusted on our development set. As an additional contribution, we defined a tuning procedure, in which we adjust our method for high precision. Following this tuning routine, our model is capable of aligning PAS pairs with a precision of 86.2%.

Heuristic Induction Method. Based on aligned pairs of PASs, the second step in our induction framework is to identify instances of implicit arguments. In Section 5, we described a computational implementation of this step in which aligned argument structures are automatically compared and discourse antecedents for implicit arguments are found by means of entity coreference chains across documents. To reduce the effect of preprocessing errors, our implementation makes use of precise preprocessing methods and a small set of restrictions that exclude instances whose automatic annotations are likely to be erroneous. We found that by combining information from different preprocessing modules, we can induce instances of implicit arguments and discourse antecedents with high precision.

Coherence Modeling and Implicit Argument Linking. To examine the utility and reliability of our data set, we additionally performed extrinsic evaluations in task-based settings. We described two particular applications of our data: linking implicit arguments to discourse antecedents and predicting coherent argument realizations. In the first application, described in Section 6, we used our data set as additional training data to enhance a pre-existing system for identifying and linking implicit arguments in discourse. Experimental results showed that the addition of our training data can improve performance in terms of precision (+15 percentage points) and F_1-score (+5 percentage points).

For the second application, we developed a novel model of local coherence, described in Section 7, which predicts whether a specific argument should be explicitly realized at a given point in discourse or whether it can already be inferred from context. Our experiments revealed that this model, when trained on automatically induced data, can predict human judgments on argument realization in newswire text with F_1-scores between 53% and 82%. In comparison, we found that entity-based coherence models from previous work only achieve results below 50%, reflecting the fact that they do not capture this phenomenon appropriately.

In conclusion, a considerable amount of work still needs to be done to enhance models for handling implicit arguments in discourse. In the long run, however, this research direction will be beneficial for many applications that involve the understanding or generation of text beyond the sentence level. In this article, we provided several research contributions that form a reliable basis for future work. In particular, we developed a framework for automatically inducing instances of implicit arguments, and we designed a novel coherence model that predicts the effect of argument realizations on perceived textual coherence. From a theoretical perspective, we validated that both explicit and implicit arguments can affect coherence and that automatically induced training data can be utilized to model this phenomenon appropriately. We further showed that our induced data set, which contains instances of implicit arguments and discourse antecedents, can be applied to enhance current models for implicit argument

linking. Future work will be able to build on these insights, further enhance existing models, and apply them to improve current state-of-the-art NLP systems.

The resources described in this article are available for download at `http://projects.cl.uni-heidelberg.de/india/`.

Acknowledgments

We thank our annotators in Heidelberg and Edinburgh. We are grateful to the anonymous reviewers for helpful feedback and suggestions. The research leading to these results has received funding by the Landes-graduiertenförderung Baden-Württemberg within the research initiative "Coherence in language processing" at Heidelberg University. Work by M. R. at the University of Edinburgh was funded by a DFG Research Fellowship (RO 4848/1-1).

References

Arnold, Jennifer E. 1998. *Reference Form and Discourse Patterns*. Ph.D. thesis, Stanford University.

Baroni, Marco and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

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

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

Belz, Anja 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, OR.

Björkelund, Anders, 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.

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

Brockett, Chris. 2007. Aligning the RTE 2006 corpus. Technical Report MSR-TR-2007-77, Microsoft Research.

Brown, Cheryl. 1983. Topic continuity in written English narrative. In Talmy Givon, editor, *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. John Benjamins, Amsterdam, pages 313–342.

Cai, Jie and Michael Strube. 2010. End-to-end coreference resolution via hypergraph partitioning. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 143–151, Beijing.

Chang, Chih-Chung and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(27):1–27.

Charniak, Eugene 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.

Chen, Desai, 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.

Chen, Zheng and Heng Ji. 2010. Graph-based clustering for computational linguistics: A survey. In *Proceedings of TextGraphs-5 - 2010 Workshop on Graph-based Methods for Natural Language Processing*, pages 1–9, Uppsala.

Cohn, Trevor, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for development and evaluation of paraphrase systems. *Computational Linguistics*, 34(4):597–614.

Das, Dipanjan, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 948–956, Los Angeles, CA.

DeNero, John, Alexandre Bouchard-Côté, and Dan Klein. 2008. Sampling alignment structure under a Bayesian translation model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 314–323, Honolulu, HI.

Di Eugenio, Barbara. 1990. Centering theory and the Italian pronominal system. In *Proceedings of the 13th Conference on Computational Linguistics*, pages 270–275, Helsinki.

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

Elsner, Micha 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, OR.

Elsner, Micha 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, OR.

Erk, Katrin and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Waikiki, HI.

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

Feng, Vanessa Wei and Graeme Hirst. 2014. Patterns of local discourse coherence as a feature for authorship attribution. *Literary and Linguistic Computing*, 29:191–198.

Filippova, Katja 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*, pages 139–142, Schloss Dagstuhl.

Fillmore, Charles J. 1986. Pragmatically controlled zero anaphora. In *Proceedings of the Twelfth Annual Meeting of the Berkeley Linguistics Society*, pages 95–107, Berkeley, CA.

Fleiss, Joseph L, Bruce Levin, and Myunghee Cho Paik. 1981. The measurement of interrater agreement. *Statistical Methods for Rates and Proportions*, 2:212–236.

Fürstenau, Hagen and Mirella Lapata. 2012. Semi-supervised semantic role labeling via structural alignment. *Computational Linguistics*, 38(1):135–171.

Gerber, Matthew and Joyce Chai. 2012. Semantic role labeling of implicit arguments for nominal predicates. *Computational Linguistics*, 38(4):755–798.

Goldberg, Andrew V. and Robert E. Tarjan. 1986. A new approach to the maximum flow problem. In *Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing*, pages 136–146, New York, NY.

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

Gorinski, Philip, 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.

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

Guo, Weiwei and Mona Diab. 2011. Semantic topic models: Combining word distributional statistics and dictionary definitions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 552–561, Edinburgh.

Hou, Yufang, Katja Markert, and Michael Strube. 2013. Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 907–917, Atlanta, GA.

Hwa, Rebecca, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.

Joshi, Aravind K. 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,* pages 435–439, Tokyo.

Kameyama, Megumi. 1985. *Zero Anaphora: The Case of Japanese*. Ph.D. thesis, Stanford University.

Karamanis, Nikiforos, Chris Mellish, Massimo Poesio, and Jon Oberlander. 2009. Evaluating centering for information ordering using corpora. *Computational Linguistics*, 35(1):29–46.

Kay, Martin and Martin Röscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19(1):121–142.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation Journal*, 42(1):21–40.

Kozhevnikov, Mikhail and Ivan Titov. 2013. Crosslingual transfer of semantic role

models. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics*, pages 1190–1200, Sofia.

Landauer, T. K. and S. T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Laparra, Egoitz 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.

Laparra, Egoitz 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.

Lee, Heeyoung, 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):885–916.

Lee, Heeyoung, 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.

Liang, Percy, Benjamin Taskar, and Dan Klein. 2006. Alignment by agreement. In *North American Association for Computational Linguistics (NAACL)*, pages 104–111, New York, NY.

Lin, Dekang. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning,* pages 296–304, Madison, WI.

Louis, Annie 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, CA.

Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, MD.

Martschat, Sebastian, 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.

McIntyre, Neil and Mirella Lapata. 2010. Plot induction and evolutionary search for story generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1562–1572, Uppsala.

Meyers, Adam. 2007. Annotation guidelines for NomBank–noun argument structure for PropBank. Technical report, New York University. Available from `http://nlp.cs.nyu.edu/meyers/nombank/nombank-specs-2007.pdf`

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

Mitchell, Jeff and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Moor, Tatjana, 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.

Padó, Sebastian and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Padó, Sebastian and Mirella Lapata. 2009. Cross-lingual annotation projection of semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340.

Palmer, Martha. 2009. Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*, pages 9–15, Pisa.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Palmer, Martha S., 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*, pages 10–19, New York, NY.

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

Pedersen, Ted. 2010. Information content measures of semantic similarity perform better without sense-tagged text. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 329–332, Los Angeles, CA.

Pitler, Emily 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, HI.

Poesio, Massimo, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.

Postolache, Oana, Dan Cristea, and Constantin Orasan. 2006. Transferring coreference chains through word alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 889–892, Genoa.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island.

Pradhan, Sameer S., Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007)*, pages 446–453, Minneapolis, MN.

Recasens, Marta and Marta Vila. 2010. Squibs: On paraphrase and coreference. *Computational Linguistics*, 36(4):639–647.

Roth, Michael 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.

Roth, Michael 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.

Ruppenhofer, Josef, Russell Lee-Goldman, Caroline Sporleder, and Roser Morante. 2012. Beyond sentence-level semantic role labeling: linking argument structures in discourse. *Language Resources and Evaluation*, 47(3):695–721.

Ruppenhofer, Josef, 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.

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

Silberer, Carina 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.

Spearman, Charles. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Sporleder, Caroline and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: A critical assessment. *Natural Language Engineering*, 14(3):369–416.

Spreyer, Kathrin and Anette Frank. 2008. Projection-based acquisition of a temporal labeller. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 489–496, Hyderabad.

Su, Fangzhong and Katja Markert. 2009. Subjectivity recognition on word senses via semi-supervised mincuts. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1–9, Boulder, C.

Tonelli, Sara 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.

Tonelli, Sara 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, OR.

Turan, Ümit Deniz. 1995. *Null vs. Overt Subjects in Turkish: A Centering Approach*. Ph.D. thesis, University of Pennsylvania.

van der Plas, Lonneke, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 299–304, Portland, OR.

Walker, Christopher, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. *ACE 2005 Multilingual Training Corpus*. Linguistic Data Consortium, Philadelphia, PA.

Wan, Stephen, Mark Dras, Robert Dale, and Cecile Paris. 2006. Using dependency-based features to take the "Para-farce" out of paraphrase. In *Proceedings of the Australasian Language Technology Workshop*, pages 131–138, Sydney.

Weischedel, Ralph, 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, PA.

Wolfe, Travis, Benjamin Van Durme, Mark Dredze, Nicholas Andrews, Charley Beller, Chris Callison-Burch, Jay DeYoung, Justin Snyder, Jonathan Weese, Tan Xu, and Xuchen Yao. 2013. PARMA: A predicate argument aligner. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 63–68, Sofia.

Yarowsky, David and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the 2nd Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 200–207, Pittsburgh, PA.

Yeh, Alexander. 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.