Casting Implicit Role Linking as an Anaphora Resolution Task

Carina Silberer[*]
School of Informatics
University of Edinburgh
Edinburgh, UK
c.silberer@ed.ac.uk

Anette Frank
Department of Computational Linguistics
Heidelberg University
Heidelberg, Germany
frank@cl.uni-heidelberg.de

Abstract

Linking implicit semantic roles is a challenging problem in discourse processing. Unlike prior work inspired by SRL, we cast this problem as an anaphora resolution task and embed it in an entity-based coreference resolution (CR) architecture. Our experiments clearly show that CR-oriented features yield strongest performance exceeding a strong baseline. We address the problem of data sparsity by applying heuristic labeling techniques, guided by the anaphoric nature of the phenomenon. We achieve performance beyond state-of-the art.

1 Introduction

A widespread phenomenon that is still poorly studied in NLP is the meaning contribution of unfilled semantic roles of predicates in discourse interpretation. Such roles, while linguistically unexpressed, can often be anaphorically bound to antecedent referents in the discourse context. Capturing such implicit semantic roles and linking them to their antecedents is a challenging problem. But it bears immense potential for establishing discourse coherence and for getting closer to the aim of true NLU.

Linking of implicit semantic roles in discourse has recently been introduced as a shared task in the SemEval 2010 competition *Linking Events and Their Participants in Discourse* (Ruppenhofer et al., 2009, 2010). The task consists in detecting unfilled semantic roles of events and determining antecedents in the discourse context that these roles can be understood to refer to. In (1), e.g., the predicate *jealousy* introduces two implicit roles, one for the experiencer, the other for the object of jealousy involved. These roles can be bound to *Watson* and the speaker (*I*) in the non-local preceding context.

(1) Watson won't allow that I know anything of art but that is mere *jealousy* because our views upon the subject differ.

(2) I_{Reader} was sitting *reading* in the chair_{Place}.

In contrast to implicit roles that can be *discourse-bound* to an antecedent as in (1), roles can be interpreted *existentially*, as in (2), with an unfilled TEXT role of the READING frame that cannot be anchored in prior discourse. The FrameNet paradigm (Fillmore et al., 2003) that was used for annotation in the SemEval task classifies these interpretation differences as definite (DNI) vs. indefinite (INI) null instantiations (NI) of roles, respectively.

2 Implicit Role Reference: A Short History

Early studies. The phenomenon of implicit role reference is not new. It has been studied in a number of early approaches. Palmer et al. (1986) treated unfilled semantic roles as special cases of anaphora and coreference resolution (CR). Resolution was guided by domain knowledge encoded in a knowledge-based system. Similarly, Whittemore et al. (1991) analyzed the resolution of unexpressed event roles as a special case of CR. A formalization in DRT was fully worked out, but automation was not addressed.

Later studies emphasize the role of implicit role reference in a frame-semantic discourse analysis. Fillmore and Baker (2001) provide an analysis of

[*] The work reported in this paper is based on a Master's Thesis conducted at Heidelberg University (Silberer, 2011).

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 1–10,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

a newspaper text that indicates the importance of frames and roles in establishing discourse coherence. Burchardt et al. (2005) offer a formalization of the involved factors: the interplay of frames and frame relations with factors of contextual contiguity. The work includes no automation, but suggests a corpus-based approach using antecedent-role coreference patterns collected from corpora.

Tetreault (2002), finally, offers an automated analysis for resolving implicit role reference. The small-scale study is embedded in a rule-based CR setup.

SemEval 2010 Task 10: Linking Roles. Triggered by the SemEval 2010 competition (Ruppenhofer et al., 2010), research on resolving implicit role reference has gained momentum again, in a field where both semantic role labeling (SRL) and coreference resolution have seen tremendous progress. However, the systems that participated in the *NI-only* task on implicit role resolution achieved moderate success in the initial subtasks: (i) recognition of implicit roles and (ii) classification as discourse-bound vs. existential interpretation (DNI vs. INI). Yet, (iii) identification of role antecedents was bluntly unsuccessful, with around 1% F-score.

Ruppenhofer et al. clearly relate the task to coreference resolution. The participating systems, though, framed the task as a special case of SRL.

Chen et al. (2010) participated with their SRL system SEMAFOR (Das et al., 2010). They cast the task as one of extended SRL, by admitting constituents from a larger context. To overcome the lack and sparsity of syntactic path features, they include lexical association and similarity scores for semantic roles and role fillers; classical SRL order and distance features are adapted to larger distances.

VENSES++ by Tonelli and Delmonte (2010) is a semantic processing system that includes lexico-semantic processing, anaphora resolution and deep semantic resolution components. Anaphora resolution is performed in a rule-based manner; pronominals are replaced with their antecedents' lexical information. For role linking, the system applies diverse heuristics including search for predicate-argument structures with compatible arguments, as well as semantic relatedness scores between potential fillers of (overt and implicit) semantic roles.

More recently Tonelli and Delmonte (2011) recur to a leaner approach for role binding, estimating a relevance score for potential antecedents from role fillers observed in training. They report an F-score of 8 points for role binding on SemEval data. However, being strongly lexicalized, their trained model seems heavily dependent on the training data.

Ruppenhofer et al. (2011) use semantic types for identifying DNI role antecedents, reporting an error reduction of 14% on Chen et al. (2010)'s results.

The poor performance results in the SemEval task clearly indicate the difficulty of resolving implicit role reference. A major factor seems to relate to data sparsity: the training set covers only 245 DNI annotations linked to an antecedent.

Linking implicit arguments of nominals. Gerber and Chai (2010) (G&C henceforth) investigate a closely related task of argument binding, tied to the linking of implicit arguments for *nominal predicates* using the PropBank role labeling scheme. In contrast to the SemEval task, which focuses on a verbs and nouns, their system is only applied to nouns and is restricted to 10 predicates with substantial training set sizes (avg: 125, median: 103).

G&C propose a discriminative model that selects an antecedent for an implicit role from an extended context window. The approach incorporates some aspects relating to CR that go beyond the SRL-oriented SemEval systems: A candidate representation includes information about all the candidates' coreferent mentions (determined by automatic CR), in particular their semantic roles (provided by gold annotations) and WordNet synsets. Patterns of semantic associations between filler candidates and implicit roles are learned for *all* mentions contained in the candidate's entity chain. They achieve an F-score of 42.3, against a baseline of 26.5.

Gerber (2011) presents an extended model that incorporates strategies suggested in Burchardt et al. (2005): using frame relations as well as coreference patterns acquired from large corpora. This model achieves an F-score of 50.3 (baseline: 28.9).

3 Casting Implicit Role Linking as an Anaphora Resolution Task

3.1 Implicit role = anaphora resolution

Recent models for role binding mainly draw on techniques from SRL, enriched with concepts from CR.

In this paper, we explicitly formulate implicit role linking as an anaphora resolution task. This is in line with the predominant conception in early work, and also highlights the close relationship with zero anaphora (Kameyama, 1985). Computational treatments of zero anaphora (e.g., Imamura et al. (2009)) are in fact employing techniques well-known from SRL. Recent work by Iida and Poesio (2011), by contrast, offers an analysis of zero anaphora in a CR architecture. Further support comes from psycholinguistic studies in Garrod and Terras (2000), who establish commonalities between implicit role reference and other types of anaphora resolution.

The contributions of our work are as follows:

i. We cast implicit role binding as a CR task, using an *entity-mention* model and discriminative classification for antecedent selection.

ii. We examine the effectiveness of model features for classical SRL vs. CR features to clarify the nature of this special phenomenon.

iii. We automatically acquire heuristically labeled data to address the sparse data problem.

i. An entity-mention model for anaphoric role resolution. In our model implicit roles that are discourse-bound (i.e. classified as DNI) are treated as anaphoric, similar to zero anaphora: the implicit role will be bound to a discourse antecedent.

In line with recent research in CR, we adopt an *entity-mention* model, where an *entity* is represented by all mentions pertaining to a coreference chain (see i.a. Rahman and Ng (2011), Cai and Strube (2010)). Our model is based on binary classifier decisions that take as input the anaphoric role and an entity candidate from the preceding discourse. The final classification of a role linking to an entity is obtained by discriminative ranking of the binary classifiers' probability estimates. Details on the system architecture are given in Section 3.2.

ii. SRL vs. CR: Analysis of feature sets. The linking of implicit semantic roles represents an interesting mixture of SRL and CR that displays exceptional characteristics of both types of phenomena.

In contrast to classical SRL, the relation between a predicate's semantic role and a candidate role filler – being realized outside the local syntactic context – cannot be characterized by syntactic path features. But similar to SRL we can compute a semantic class type expected by the role and determine which candidate is most appropriate to fill the semantic role.

Anaphoric binding of unfilled roles also diverges from classical CR in that the anaphoric element is not overtly expressed. This excludes typical CR features that refer to overt realization, such as agreement or string overlap. Again, we can make use of a semantic characterization of role fillers to determine the role's most appropriate antecedent entity in the discourse. This closely relates to semantic class features employed in CR (e.g., Rahman and Ng (2011)).

Thus, semantic association features are important modeling aspects, but they do not contribute to clarifying the nature of the phenomenon. We will include additional properties that are considered characteristic for CR, such as the semantics of an *entity* (as opposed to individual mentions), or salience properties of antecedents (cf. Section 4.3). Thus, the model we propose substantially differs from prior work.

We classify the features of our models as SRL vs. CR features, plus a mixture class that relates to both phenomena. We examine which type of features is most effective for resolving implicit role reference.

iii. Heuristic data acquisition. In response to the sparse data problem encountered with the SemEval data set and the general lack of annotated resources for implicit role binding, we experiment with techniques for heuristic data acquisition. The strategy we apply builds on our working hypothesis that implicit role reference is best understood as a special case of (zero) anaphora resolution.

We process manually annotated coreference data sets that are jointly labeled with semantic roles. From these we extract entity chains that contain anaphoric pronouns that fill a predicate's semantic role. We artificially delete the pronoun's role label and transfer it to its closest antecedent in its chain. In this way, we convert the example to an instance that is structurally similar to one involving a locally unfilled semantic role that is bound to an overt antecedent. An example is given below: in (3.a) we identify a pronoun that fills the SPEAKER role of the frame STATEMENT. We transfer this role label to its closest antecedent (3.b).

(3) a. Riady$_k$ spoke in his$_k$ 21-story office building on the outskirts of Jakarta. [...] The timing of $\underline{\text{his}}_{k,Speaker}$ $\underline{\text{statement}}_{Statement}$ is important.

b. Riady$_k$ spoke in $\underline{\text{his}}_{k,Speaker}$ 21-story office building on the outskirts of Jakarta. [...] The timing of $\emptyset$ $\underline{\text{statement}}_{Statement}$ is important.

Clearly such artificially created annotation instances are only approximations of naturally occurring cases of implicit role binding. But we expect to acquire numerous data points for relevant features: semantic class information for the antecedent entity, the predicate's frame and roles and coherence properties.

3.2 System Architecture

Our approach is embedded in an architecture for supervised CR using an entity-mention model. The main processing steps of the system include: (1) entity detection, (2) instance creation with feature extraction and (3) classification. As we are focusing on the resolution of implicit DNI roles, we assume that the text is already augmented with standard CR information (we make use of gold data and automatically assigned coreference chains). Accordingly, the description of modules focuses exclusively on the resolution of DNIs.

(1) Entity Detection. We first collect the entire entity set $\mathcal{E}$ mentioned in the discourse. This set forms the overall set of candidates to consider for DNI linking. For each DNI d_k to be linked, a subset of candidates $\mathcal{E}_k \subset \mathcal{E}$ is chosen as candidate search space for resolving d_k. We experiment with different strategies for constructing $\mathcal{E}_k$ (cf. Section 4).

(2) Instance Creation. The next step consists in the creation of (training) instances for classification including the extraction of features for all instances.

An instance $inst_{e_j,d_k}$ consists of the active DNI d_k, its frame and a candidate entity $e_j \in \mathcal{E}_k$. Instance creation follows an entity-based adaption of the standard procedure of Soon et al. (2001), which has been applied by Yang et al. (2004, 2008). Processing the discourse from left to right, for each DNI d_k, instances $\mathcal{I}_k$ are created by processing $\mathcal{E}_k$ from right to left according to each entity's most recent mention, starting with the entity closest to d_k. Note that, as entities instead of mentions are considered, only one instance is created for an entity which is mentioned several times in the search space.

In training, the instance creation stops when the correct antecedent, i.e. a positive instance, as well as at least one negative instance have been found.[1]

(3) Classification. From the acquired training instances we learn a binary classifier that predicts for an instance $inst_{e_j,d_k}$ whether it is positive, i.e. entity e_j is a correct antecedent for DNI d_k. Further, the classifier provides a probability estimate for $inst_{e_j,d_k}$ being positive. We obtain classifications for all instances in $\mathcal{I}_k$. Among the positive classified instances, we select the antecedent e with the highest estimate. That is, we apply the *best-first* strategy (Ng and Cardie, 2002). In case of a tie, we choose the antecedent which is closer to the target. If no instance is classified as positive, d_k is left unfilled.

4 Data and Experiments

4.1 SEMEVAL 2010 task and data set

We adhere to the SemEval 2010 task by Ruppenhofer et al. (2009) as test bed for our experiments. The main focus of our work is on part (iii), the identification of antecedents for DNIs. Subtasks (i) and (ii), the recognition and interpretation of NIs will be only tackled to enable comparison to the participating systems of the SemEval *NI-only* task.

The SemEval task is based on fiction stories by A. C. Doyle, one story as training data and another two chapters as test set, enriched with coreference and FrameNet-style frame annotations. Information about the training section is found in Table 1. The test data comprise 710 NIs (349 DNIs, 361 INIs), of which 259 DNIs are linked.

4.2 Heuristic data acquisition

Since the training data has a critically small amount of linked DNIs, we heuristically labeled training data on the basis of data sets with manually annotated coreference information: OntoNotes 3.0 (Hovy et al., 2006), as well as ACE-2 (Mitchell et al., 2003) and MUC-6 (Chinchor and Sundheim, 2003).

OntoNotes 3.0 was merged with gold SRL annotations from the CoNLL-2005 shared task. By means of SemLink-1.1 (Loper et al., 2007) and a mapping included in the SemEval data, these PropBank (PB, Palmer et al. (2005)) annotations were

[1] We additionally impose several restrictions, e.g., a valid candidate must not already fill another role of the active frame.

	#ent	avg	avg	#frames	#frame	#DNI	#DNI
		#ent/doc	size		types		types
SemEval	141	141	9	1,370	317	245	155
ONotes	7899	23	3	12,770	258	2,220	270
ACE-2	3564	11	4	58,204	757	4,265	578
MUC-6	1841	15	3	20,140	654	997	310

corpus	coref	semantic roles
ONotes	manual	manual PB CoNLL05, ported to FN
ACE-2	manual	automatic FN (Semafor)
MUC-6	manual	automatic FN (Semafor)

Table 1: SemEval vs. heuristically acquired data

mapped to their FrameNet (FN) counterparts, if existent. For the ACE-2 and MUC-6 corpora, we used Semafor (Das and Smith, 2011) for automatic annotation with FN semantic roles. From these data sets we acquired heuristically annotated instances of role linking using the strategy explained in 3.1.

Table 1 summarizes the resulting training data. The heuristically labeled data extends the manually labeled DNI instances by an order of magnitude.

4.3 Model parameters

Entity sets $\mathcal{E}_{dni}$. For definition of the set of candidate entities to consider for DNI linking, $\mathcal{E}_{dni}$, we determined different parameter settings with restrictions on the types, distances and prominence of candidate antecedents. For instance, unlike in noun phrase CR, antecedents for a DNI can be realized by a wide range of constituents other than NPs, such as prepositional (`PP`), adverbial (`ADVP`), verb phrases (`VP`) and even sentences (`S`) referring to propositions.

These settings, stated in Table 2, were inferred by experiments on the training data and by examining its statistics: *AllChains* is motivated by the fact that 72% of the DNIs are linked to referents with non-singleton chains. On the other hand, the majority of DNI antecedents – not only non-singletons, but also phrases of a certain type or terminals that overtly fill other roles – are located in the current and the two preceding sentences (69.6%), which motivates *SentWin*. However, antecedents are also located far beyond this window span which is probably due to the nature of the SemEval texts, with prominent entities being accessible over longer stretches of discourse. *Chains+Win* is designed by taking into account all previous observations.

AllChains This set contains all the entities represented by non-singleton coreference chains that were introduced in the discourse up to the current DNI position, assuming that this way only more salient entities are considered.

SentWin Comprises constituents with a certain phrase type[2] or terminals that overtly fill a role, occurring within the current or the preceding two sentences.

Chain+Win This set comprises **SentWin** plus all entities mentioned at least five times up to the current DNI position (i.e. salient entities).

Table 2: Entity set settings $\mathcal{E}_{dni}$

Training data sets. We made use of different mixtures of training data: SemEval plus different extensions using the heuristically acquired data summarized in Table 1.

4.4 Feature sets: SRL, mixed and CR-oriented

Table 3 lists the most important features used for training our models. Features 1-13 were used in the best model and are ordered by their strength based on feature ablation experiments (cf. Section 5). All features are marked for their general type; the last column marks features employed by G&C.[3]

Below we give some details for selected features.

Feat. 1: Prominence. We first compute average prominence of an entity e (Eq. 2) by summing over the size (= nb. of mentions) of all entities e in a window w[4] of preceding sentences and dividing by the nb. of entities E in w. Prominence of e (Eq. 1) is set to the difference between its size in w and the average prominence score.[5] The final feature value records the relative rank of e's prominence score compared to the scores of the other candidates.

$$prom(e, w) = \#mentions(e, w) - avg\ prom(w) \quad (1)$$

$$avg\ prom(w) = \frac{\sum_{e \in E} \#mentions(e, w)}{|E|} \quad (2)$$

[2] The phrase type must be `NPB`, `S`, `VP`, `SBAR`, or `SG`.

[3] $\sim$ marks features that are similar to G&C features. Note that their only CR features are distance features.

[4] We set $w = 2$ based on experiments on the training data.

[5] This prominence score was proposed by Dolata (2010) within an entity grid approach to role linking.

nr	feature		type	G&C
1	prominence	prominence score of the entity in the current discourse position	CR	-
2	pos.dist_mention	PoS or phrase type of the most recent explicit mention concatenated with sentence distance to the target	(CR)	-
3	dist_mentions	minimum distance between DNI and entity in mentions	CR	-
4	dist_sentences	minimum distance between DNI and entity in sentences	CR	+
5	vnroles_dni.entity	the counterparts of the DNI in VerbNet (VN, Kipper et al. (2000)) concatenated with the VN roles the entity already instantiates	mixed	+
6	roles_dni.entity	concatenation of the DNI with the FN roles the entity already instantiates	mixed	$\sim$
7	semType_dni.entity	semantic type of the DNI concatenated with the semantic types of the roles the entity already instantiates	mixed	-
8	avgDist_sentences	average sentence distance between the entity and the DNI	CR	+
9	sp_supersense	agreement of the selectional preferences for the DNI and the most frequent supersense of the entity	mixed	-
10	function (target)	grammatical function of the target	SRL	-
11	wnss_ent.st_dni	pointwise mutual information between the entity's WN supersense ss and the DNI's FN semantic type st: $pmi(ss, st) = log_2 P(ss|st)/P(ss)$	mixed	-
12	nbRoles_dni.entity	like feature 5, but with NomBank arguments 0 and 1	mixed	$\sim$
13	frame.dni	frame name concatenated with the DNI	SRL	-

Table 3: Best features used for training. Feat. 11 was computed on the FN dataset and the SemEval training data.

Feat. 9: SelPrefs. We compute selectional preferences following the information-theoretic approach of Resnik (1993, 1996). Similar to Erk (2007), we used an adapted version which we computed for semantic roles by means of the FN database rather than for verb argument positions. The WordNet classes over which the preferences are defined are WordNet lexicographer's files (supersenses).

The selectional association values $\Lambda(dni, ss)$ of the DNI's selectional preferences are retrieved for the supersense ss of each candidate antecedent's head. As for Feat. 1, we define a candidate's feature value by its rank in the ordered list of these Λs.

4.5 Experiments

Evaluation measures. We adopt the precision (P), recall (R) and F_1 measures in Ruppenhofer et al. (2010). A true positive is a DNI which has been linked to the correct entity as given by the gold data.

Classifiers and feature selection. For DNI linking, we use BayesNet (Cooper and Herskovits, 1992) as classifier, implemented in Weka (Witten and Frank, 2000).[6] For each parameter combination, we perform feature selection by means of leave-one-out 10-fold cross-validation on the SemEval training data with successively removing/determining the best features. The resulting models M_i are then evaluated on the SemEval test data in different setups:

Exp1: Linking DNIs. Exp1 evaluates our models on the DNI linking task proper (NI-only step (iii)). This setting uses the gold coreference, SRL and DNI information in the test data.

Exp2: Full NI-only. For benchmarking on the SemEval task, we perform the complete NI-only task. Here, the test data is only enriched w/ SRL labeling. Each frame *f* in the test corpus is processed, involving the following steps:

(i) *Recognition of NIs* is performed by consulting the FN database[7] and determining the FN core roles that are unfilled. From this NI set, roles that are conceptually redundant or competing with *f*'s overt roles are rejected as they don't need to or must not be linked, respectively.

(ii) For predicting the *interpretation of an NI*, we use LibSVM (Chang and Lin, 2001) as classifier which further assigns each NI a probability estimate of the NI being definite. We use a small set of features: the FN semantic type of the NI and a boolean feature indicating whether the target is in passive voice and the agent (object) not realized. Further, we use a statistical feature which gives the relative

[6] We experimented with different learners and selected the algorithm that performed best for the different subtasks.

[7] We used the FrameNetAPI by Reiter (2010).

model	add. data	entity set	frame anno.	**DNI Linking** (%) P	 R	 F_1
M_0	-	AllChains	gold	25.6	25.1	25.3
M_1	ON2-10	Chains+Win	proj	30.8	**25.1**	**27.7**
$\mathrm{M}_{1'}$	ON2-24	AllChains	proj	**35.6**	20.1	25.7
$\mathrm{M}_{1''}$	ON2-24	SentWin	proj	23.3	22.4	22.8
M_2	MUC	Chains+Win	auto	26.1	24.3	25.3
M_3	ACE	AllChains	auto	24.0	21.2	22.5
Prom	–	Chains+Win	–	20.5	20.5	20.5

Table 4: Exp1: Best performing models for different entity and data settings. Test data contain gold CR chains.

frequency of the role's realization as DNI and INI, respectively, in the training data.

(iii) *DNI linking* is performed for each of *f*'s predicted DNIs $\mathcal{D}_f$ in descending order of their probability estimates. If an antecedent e_m can be determined for a predicted DNI, the role is labeled as such and linked to e_m. As the DNI's role has been filled now, competing or redundant DNIs are removed from $\mathcal{D}_f$ before moving to the next predicted DNI. Only DNIs for which an antecedent is found are labeled as such.

Exp2 is evaluated on both gold coreference annotation and automatically assigned coreference chains, using the CR system of Cai et al. (2011).

5 Evaluation and Results

5.1 Exp1: DNI linking evaluation

Table 4 shows the best performing models for DNI linking for each parameter setting[8]. We compare them to a strong baseline *Prom* (last row) that links each DNI to the antecedent candidate with highest prominence score. Its F_1-score is beaten by the other models, with a gain of 7.2 points for model M_1. The high performance of the baseline can be taken as evidence that salience factors are crucial for this task.

The best performing model M_1 (27.7 F_1) uses about a fifth of the ON data with *Chains+Win*. When using *SentWin* as entity set, F_1 drops to 18.5 (not shown). The best performing model using *SentWin* ($M_{1''}$) performs 4.9 points below M_1. Hence, reliance on the *Chains+Win* set seems beneficial. Performance of the *AllChains* setting varies over the different data sets: the strongest model is M_0 without additional data. An explanation could be the different data domains (story vs. news), leading to a different nature (length and number) of the entities.

Features	**P** (%)	**R** (%)	$\mathbf{F}_1$ (%)
all	30.8	25.1	27.7
- 1-4,8 (CR)	21.6	8.1	11.8
- 10,13 (SRL)	31.0	25.9	28.2
- 5-7,9,11-12 (mixed)	20.6	20.5	20.5

Table 5: Results of ablation study.

In general, the models seem to profit from heuristically labeled training data. We note strong gains (up to 10 pts) in precision for 3 of these 5 best models, compared to M_0. Finally, we observe higher performance when using additional data with gold/ projected semantic frame annotations (M_1, $M_{1'}$).

Analysis of the best model. Table 5 states the results for M_1 when leaving out one of the feature types at a time. The serious drop of F_1 from 27.7% to 11.8% when omitting CR features clearly demonstrates that this feature type has by far the greatest impact on the task performance. Rejection of the mixed features decreases F_1 to a score equal to the prominence baseline, whereas leaving out the SRL-features even slightly increases F_1. The weakness of Feature 13 could still be attributed to data sparsity.

5.2 Exp2: Full NI-only evaluation

Table 6 lists the results for the full NI-only task obtained with the presented models with different additional training data sets (lines 2-5). When performing all three steps, the F_1-score of the best model M_1 drops to 10.1% (-17.6 pts, col. 10) under usage of automatic coreference annotations in the test data (i.e. under the real task conditions). When using gold coreference annotations, the F_1-score is at 18.1% (col. 11), which can be seen as an upper bound for our current models on this task. The difference of 9.6 points between only performing DNI linking (Table 4) and the full NI-only task reflects the fact that recognizing (step i) and interpreting (step ii) NIs bear difficulties on their own.[9]

Comparison of our models with the two SemEval

[8] We consider the 3 types of entity sets and different training setups ± additional data (Section 4.3); additional data with gold, projected or automatic frame annotations. The ON data was also evaluated with roughly a fifth of ON to evaluate the effect of different amounts of data of the same type of data.

[9] When not performing step (iii), NI recognition achieves 77.6% recall and 67% relative precision.

				Null Instantiations (%)						
model	**add.**	**entity**	**frame**	**recogn.**	**interpret.** (precision)		**DNI Linking** (%)			
	data	**set**	**anno.**	recall	relative	absolute	P	R	F_1	F_1(crf)
M_0	-	AllChains	gold	58	68	40	6.0	8.9	7.1	12.5
M_1	ON2-10	Chains+Win	proj	56	69	38	9.2	11.2	10.1	18.1
M_2	MUC	Chains+Win	auto	52	70	36	7.0	8.5	7.6	11.0
M_3	ACE	AllChains	auto	56	68	38	5.9	8.1	6.8	11.3
$M_{3'}$	ACE	Chains+Win	auto	56	68	38	6.9	9.7	8.0	9.5
SEMAFOR		–		63	55	35			1.40	
VENSES++		–		8	64	5			1.21	
T&D		–		54	75	40	13.0	6.0	8	

Table 6: Exp2 results obtained for our models (lines 1-5) and comparable systems (lines 6-8). Column 5 gives the score for correctly recognized NIs. Cols. 6 and 7 report precision for correctly interpreted NIs on the basis of the correctly recognized (relative) vs. all gold NIs to be recognized (absolute). The scores in the last column (F_1(crf)) were obtained with gold CR annotations.

task participants[10] (lines 7-8) shows that our models clearly outperform these systems – with a gain of +5.7 and +8.89 points in F_1-score in DNI linking.[11]

Compared to Tonelli and Delmonte (2011) (T&D), M_1 has a higher F_1-score in linking of +2.1 points. In contrast to our method, their linking approach is (admittedly) heavily lexicalized and strongly tailored to the domain of the used data.

6 Conclusion

We cast the problem of linking implicit semantic roles as a special case of (zero) anaphora resolution, drawing on insights from earlier work and parallels observed with zero anaphora. Our results strongly support this analysis: (i) Feature selection clearly determines CR-related features as strongest support for DNI linking. (ii) Our models beat a strong baseline using a prominence score to determine DNI reference. (iii) We devise a method for heuristically labeling training data that simulates implicit role reference. Using this data we obtain system performance beyond state-of-the-art, with high gains in precision.

While these findings clearly corroborate our conceptual approach, overall performance is still meager. Comparison to G&C's setting suggests that training data is a serious issue. We addressed the problem of training set size using heuristic data acquisition. The nature of semantic role annotations may be another problem, as FrameNet-style roles do not generalize well. Finally, implicit roles pertaining to nominalizations tend to be more local than those pertaining to verbs[12] and might be less diverse.

Our model is closer in spirit to G&C than the SemEval systems, but differs by being embedded in an entity-based CR architecture using discriminative antecedent selection. Also, we address a more principled issue, by exploring the nature of the task using a qualitative feature analysis. Our system compares favorably to related work. Benchmarking against the SemEval participants and T&D shows clear improvements. Also, T&D's model is closely tied to domain data, while ours is enhanced with out-of-domain data. Exact comparison to G&C needs to be conducted on the same data set and labeling scheme.

In sum, within the chosen setting we can show that implicit role reference is best modeled as a special case of anaphora resolution. We observe that models trained on cleaner data perform better than on larger, but more noisy data sets. Thus, it is essential to further enhance the quality of heuristically labeled data. Applying the classifiers for steps (i) and (ii) as a filter could help to better constrain the data to the target phenomenon.

Acknowledgements. We would like to thank Mateusz Dolata for his help with salience and coherence features, and Michael Roth for his server support.

[10] The F_1-scores are from `http://semeval2.fbk.eu/semeval2.php?location=Rankings/ranking10.html`

[11] Moreover, note that Ruppenhofer et al. describe a weaker evaluation, that judges DNI linkings as correct if the span of the linked referent contains the gold referent. Further, they consider 14 linked INIs in the test data, although linking INIs conflicts with the definition of INIs.

[12] This is confirmed by analysis of the SemEval vs. NomBank corpus of G&C.

References

Aljoscha Burchardt, Anette Frank, and Manfred Pinkal. 2005. Building Text Meaning Representations from Contextually Related Frames – A Case Study. In *Proceedings of the 6th International Workshop on Computational Semantics*, IWCS-6, pages 66–77, Tilburg, The Netherlands.

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

Jie Cai, Eva Mújdricza-Maydt, and Michael Strube. 2011. Unrestricted coreference resolution via global hypergraph partitioning. In *Proceedings of the Shared Task of 15th Conference on Computational Natural Language Learning*, pages 56–60, Portland, Oregon.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a Library for Support Vector Machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Desai Chen, Nathan Schneider, Dipanjan Das, and Noah A. Smith. 2010. SEMAFOR: Frame Argument Resolution with Log-Linear Models. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 264–267, Uppsala, Sweden, July.

Nancy Chinchor and Beth Sundheim, 2003. *Message Understanding Conference (MUC) 6*. Linguistic Data Consortium, Philadelphia.

Gregory F. Cooper and Edward Herskovits. 1992. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, 9(4):309–347.

Dipanjan Das and Noah A. Smith. 2011. Semi-supervised frame-semantic parsing for unknown predicates. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *ACL*, pages 1435–1444. The Association for Computer Linguistics.

Dipanjan Das, 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, California, June.

Mateusz Dolata. 2010. *Extending the Entity-Grid Model for the Processing of Implicit Roles in Discourse*. Bachelor's thesis, Department of Computational Linguistics, Heidelberg University, Germany.

Katrin Erk. 2007. A Simple, Similarity-based Model for Selectional Preferences. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, ACL '07, pages 216–223, Prague, Czech Republic, June.

Charles J. Fillmore and Collin F. Baker. 2001. Frame Semantics for Text Understanding. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, June.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R. L. Petruck. 2003. Background to Framenet. *International Journal of Lexicography*, 16(3):235–250.

Simon Garrod and Melody Terras. 2000. The Contribution of Lexical and Situational Knowledge to Resolving Discourse Roles: Bonding and Resolution. *Journal of Memory and Language*, 42(4):526–544.

Matthew Gerber and Joyce Chai. 2010. Beyond NomBank: A Study of Implicit Arguments for Nominal Predicates. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1583–1592, Uppsala, Sweden, July.

Matthew Steven Gerber. 2011. *Semantic Role Labeling of Implicit Arguments for Nominal Predicates*. Ph.D. thesis, Michigan State University.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, HLT-NAACL '06, pages 57–60, New York, New York, June.

Ryu Iida and Massimo Poesio. 2011. A Cross-Lingual ILP Solution to Zero Anaphora Resolution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 804–813, Portland, Oregon.

Kenji Imamura, Kuniko Saito, and Tomoko Izumi. 2009. Discriminative Approach to Predicate-Argument Structure Analysis with Zero-Anaphora Resolution. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, ACL-IJCNLP '09, pages 85–88, Suntec, Singapore, August.

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

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-Based Construction of a Verb Lexicon. In *Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, pages 691–696, Austin, Texas. AAAI Press. `http://verbs.colorado.edu/~mpalmer/projects/verbnet.html`.

Edward Loper, Szu ting Yi, and Martha Palmer. 2007. Combining Lexical Resources: Mapping between

PropBank and VerbNet. In *Proceedings of the 7th International Workshop on Computational Linguistics*.

Alexis Mitchell, Stephanie Strassel, Mark Przybocki, JK Davis, George Doddington, Ralph Grishman, Adam Meyers, Ada Brunstein, Lisa Ferro, and Beth Sundheim, 2003. *ACE-2 Version 1.0*. Linguistic Data Consortium, Philadelphia.

Vincent Ng and Claire Cardie. 2002. Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 104–111, Philadelphia, Pennsylvania.

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

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, March.

Altaf Rahman and Vincent Ng. 2011. Narrowing the modeling gap: A cluster-ranking approach to coreference resolution. *Journal of Artificial Intelligence Research*, 40:469–521.

Nils Reiter. 2010. FrameNet API. `http://www.cl.uni-heidelberg.de/trac/FrameNetAPI`.

Philip Resnik. 1996. Selectional Constraints: an Information-theoretic Model and its Computational Realization. *Cognition*, 61(1-2):127–159, November.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2009. SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of the NAACL-HLT 2009 Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-09)*, pages 106–111, Boulder, Colorado, June.

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

Josef Ruppenhofer, Philip Gorinski, and Caroline Sporleder. 2011. In Search of Missing Arguments: A Linguistic Approach. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 331–338, Hissar, Bulgaria, September.

Carina Silberer. 2011. *Linking Implicit Semantic Roles in Discourese Using Coreference Resolution Methods*. Master's thesis, Department of Computational Linguistics, Heidelberg University, Germany.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27:521–544, December.

Joel R. Tetreault. 2002. Implicit Role Reference. In *International Symposium on Reference Resolution for Natural Language Processing*, pages 109–115, Alicante, Spain.

Sara Tonelli and Rodolfo Delmonte. 2010. VENSES++: Adapting a Deep Semantic Processing System to the Identification of Null Instantiations. In *Proceedings of the 5th International Workshop on Semantic Evaluations*, pages 296–299, Uppsala, Sweden, July.

Sara Tonelli and Rodolfo Delmonte. 2011. Desperately Seeking Implicit Arguments in Text. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, pages 54–62, Portland, Oregon, USA, June.

G. Whittemore, M. Macpherson, and G. Carlson. 1991. Event-building through role filling and anaphora resolution. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, pages 17–24, Morristown, NJ, USA.

Ian H. Witten and Eibe Frank. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA, USA.

Xiaofeng Yang, Jian Su, Guodong Zhou, and Chew Lim Tan. 2004. An NP-cluster Based Approach to Coreference Resolution. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, pages 226–232, Geneva, Switzerland.

Xiaofeng Yang, Jian Su, Jun Lang, Chew Lim Tan, Ting Liu, and Sheng Li. 2008. An Entity-Mention Model for Coreference Resolution with Inductive Logic Programming. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, ACL '08:HLT, pages 843–851, Columbus, Ohio, June.