Sources of Evidence for Implicit Argument Resolution

Egoitz Laparra
IXA Group
Basque Country University
San Sebastian, Spain
egoitz.laparra@ehu.es

German Rigau
IXA Group
Basque Country University
San Sebastian, Spain
german.rigau@ehu.es

Abstract

Traditionally, semantic role labelling systems have focused on searching the fillers of those explicit roles appearing within sentence boundaries. However, when the participants of a predicate are implicit and can not be found inside sentence boundaries, this approach obtains incomplete predicative structures with null arguments. Previous research facing this task have coincided in identifying the implicit argument filling as a special case of anaphora or coreference resolution. In this work, we review a number of theories that model the behaviour of discourse coreference and propose some adaptations to capture evidence for the implicit argument resolution task. We empirically demonstrate that exploiting such evidence our system outperforms previous approaches evaluated on the SemEval-2010 task 10 dataset. We complete our study identifying those cases that traditional coreference theories can not cover.

1 Introduction

One of the most relevant tasks in the semantic processing of texts is identifying the arguments of a predicate. Several systems have been developed to perform this task, called Semantic Role Labelling (SRL) (Gildea and Jurafsky, 2000). However they have traditionally focused on searching the fillers for the overtly realized arguments in the local context of the predicate. In other words, only exploring those participants that share a syntactical relation with the predicate. Since traditional SRL systems depend strongly on these syntactic relations, they cannot perform predictions when the candidate instantiation of the argument is not explicit. Nevertheless, *some* null instantiated arguments can be inferred from the context. Using the nominal predicates of NomBank (Meyers et al., 2004), Gerber and Chai (Gerber and Chai, 2010) pointed out that the implicit arguments can add up to 65% to the coverage of the instantiations. As a consequence, increasing the number of connections between the predicates and their participants could help dramatically text understanding.

In FrameNet (Baker et al., 1998), the predicates, called *lexical-units* (LU), evoke frames which roughly correspond to different events or scenarios. For each frame, a set of possible arguments are defined. These arguments are called *Frame Elements* (FE) and when they are not explicitly instantiated they are called Null Instantiations (NI). When they can be inferred implicitly they are called *Definite Null Instantiations* (DNI). In the next example, the LU **tenant**$_n$ evoking the frame **Residence** has an instantiated FE, *Resident*, whose filler is [the tenants]. The correct filler for the DNI corresponding to FE *Location*, [*the house*], appears two sentences before:

> "Now, Mr. Holmes, with your permission, I will show you round the house." The various bedrooms and sitting-rooms had yielded nothing to a careful search. Apparently [the **tenants**$_{Residence}$]$_{Resident}$ had brought little or nothing with them. DNI$_{Location}$

Early studies on implicit arguments described this problem as a special case of anaphora or coreference resolution (Palmer et al., 1986; Whittemore et al., 1991; Tetreault, 2002). Also recent works cast this problem as an anaphora resolution task (Silberer and Frank, 2012).

In this work we present a detailed study of a set of features that have been traditionally used to model anaphora and coreference resolution tasks. We describe how these features manifest in a FrameNet based corpus for modeling implicit argument resolution, including an analysis of their benefits and drawbacks.

The paper is structured as follows: section 2 discusses the related work. Section 3 describes the SemEval-2010 task 10 dataset. Section 4 reviews a number of sources of evidence applied to the anaphora or coreference resolution tasks. We also propose how to adapt these features to select the appropriate fillers for the implicit arguments. Section 5 presents some experiments we have carried out to test these features. Section 6 discusses the initial results. Finally, section 7 offers some concluding remarks and presents some future researching.

2 Related Work

Task 10 of SemEval-2010 focused on the evaluation of SRL systems based on the FrameNet paradigm[1] (Ruppenhofer et al., 2009). This task was divided in two different sub-tasks:

(i) Argument annotation in a traditional SRL manner.
(ii) Filling null instantiations over the document.

The systems participating in the second subtask identified those missing *Frame Elements* that were really *Null Instantiations*, decided which of those NI were definite, and finally located the correct fillers of the DNIs. Two systems participated in the second sub-task: VENSES++ and SEMAFOR.

VENSES++ (Tonelli and Delmonte, 2010) builds logical rules from syntactic parsing and uses hand-crafted lexicons. They apply a rule based anaphora resolution procedure before employing semantic similarity between a NI and a potential filler using WordNet (Fellbaum, 1998). More recently, the same authors have tried to improve the performance of their system (Tonelli and Delmonte, 2011).

SEMAFOR (Chen et al., 2010) is a supervised system that extends an existing semantic role labeller replacing the features defined for regular arguments with two new semantic features. First, their system checks if a potential filler in the context fills the null-instantiated role in one of the FrameNet sentences, and second, it calculates the distributional semantic similarity between the fillers and the roles. Although this system obtained the best performance in the task, data sparseness strongly affected the results.

In a different approach, (Ruppenhofer et al., 2011) explore a number of linguistic strategies in order to enhance the DNI identification. They conclude that a more sophisticated approach for DNI identification can improve significantly the performance of the whole pipeline, even if the method for the DNI filling is simple. For filling DNIs they propose to use the semantic types specified for FEs in FrameNet. Following this line (Laparra and Rigau, 2012) presented a novel strategy for the DNI identification exploiting explicit Frame Elements annotations. Their approach gets the best results in the state of the art for DNI identification and showed its relevance in the DNI filling process.

(Silberer and Frank, 2012) propose to solve the task adapting an entity-based coreference resolution model. In this work, the authors also extend automatically the training corpus to avoid data sparseness.

Finally, (Gerber and Chai, 2010) define a closely related task characterizing the implicit arguments of some predicates appearing in NomBank (Meyers et al., 2004). They use a set of syntactic and semantic features to train a logistic regression classifier. The documents, obtained from the Wall Street Journal corpus, were already annotated with explicit arguments. Unlike SemEval-2010 task, the resulting dataset contains 1,253 predicate instances with an average of 1.8 roles annotated per instance. However just a set of ten different predicates is taken into account.

3 SEMEVAL-2010 dataset

In the experiments reported in this paper, we have used the dataset distributed in SemEval-2010 for Task 10 "Linking Events and their Participants in Discourse". The corpus contains some chapters extracted

[1] http://www.coli.uni-saarland.de/projects/semeval2010_FG/

from two Arthur Conan Doyle's stories. "The Tiger of San Pedro" chapter from "The Adventure of Wisteria Lodge" was selected for training, while chapters 13 and 14 from "The Hound of the Baskervilles" were selected for testing. The texts are annotated using the frame-semantic structure of FrameNet 1.3 including null instantiations, the type of the NI and the corresponding fillers for each DNI. Table 1 shows the number of DNI in the dataset.

data-set	DNIs (solved)	Explicit FE
train	303 (245)	2,726
test-13	158 (121)	1,545
test-14	191 (138)	1,688

Table 1: Number of DNI and Explicit FE annotations for the SemEval-10 Task-10 corpus.

The dataset also includes the annotation files for the lexical units and the full-text annotated corpus from FrameNet. The annotations are enriched with a constituent-based parsing and for the training document there are manual coreference annotations available.

4 Sources of evidence

Many sources of evidence have proved their utility in reference resolution (Burger and Connolly, 1992). In this section, we adapt them to the specific characteristics of the DNI linking task. We also present their behaviour over the training data. Two main differences must been taken into account with respect to anaphora and coreference tasks. First, in anaphora and coreference tasks, mentions occur explicitly and they can be exploited to check particular constrains. Without an explicit argument, in some cases, we decided to obtain the evidences from the predicate (that is, the lexical-unit) of the target DNI. Second, the referenced entities are not just nouns or pronouns but also verbs, adjectives, etc. Therefore, some features must be generalized. We introduce the sources of evidence grouped in four different types.

4.1 Syntactic

Some of the earliest theories studying pronoun resolution focused on the syntactic relations between the referenced entities. Here we present two syntactic features that also exploit this source of evidence. In both cases, we also include an artificial node covering all document sentence trees in order to generalize its behaviour beyond sentence boundaries.

Command: C-command (Reinhart, 1976) is a syntactic relationship between nodes in a constituency tree. One node N1 is said c-commanded by another N2 if three requirements are satisfied:

- N1 does not dominate N2
- N2 does not dominate N1
- The first upper node that dominates N1, also dominates N2

This syntactic relation has proved to be useful to locate anaphoric references. Now, we study if this relationship can also be of utility for DNI resolution. We implemented this relation as a distance measure between the candidate filler node and the nearest common ancestor with respect the lexical-unit of the target DNI (see a simple example in figure 1). Note that a value equal to zero means that either the filler dominates the target or the target dominates the filler. Besides, those fillers having a command value equal to one satisfy the c-command theory. Figure 2 presents the frequency distribution of our distance measure on the training data. It seems that most fillers have a command value equal or close to one.

Nearness: The constituency tree can also be exploited for anaphora resolution using breadth-first search techniques. A widely known algorithm based in this search is the Hobbs' algorithm (Hobbs, 1977). This algorithm follows a traversal search of the tree looking for a node that satisfies some constraints. Because of the nature of these constraints this algorithm cannot be directly applied to the implicit argument annotation task. Instead, we studied if the breadth distance can be an evidence through a measure we call **nearness**. We calculate **nearness** N as follows:

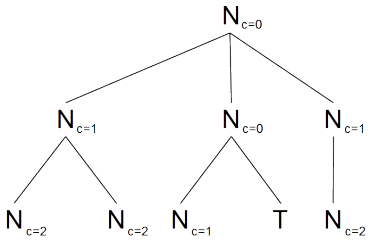

Figure 1: Sample values of **command** for different nodes in a constituency tree. T represents the lexical-unit of the target DNI

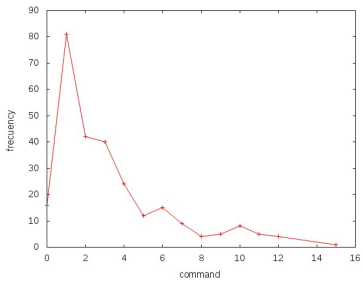

Figure 2: Frequency distribution of the different values of **command** in the training data

- P is the first upper node that dominates the lexical-unit T and the filler F
- B is the tree branch containing F whose parent is P
- If F precedes T, N is the number of following siblings of F in B
- If F follows T, N is the number of preceding siblings of F in B
- If T dominates F or F dominates T, N is equal to 0

Figure 3 presents some examples of values obtained using this measure. Figure 4 shows the frequency distribution of the different values of **nearness** in the training data. It also seems that most fillers prefer small **nearness** values.

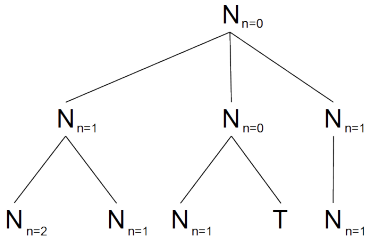

Figure 3: Sample values of **nearness** for different nodes in a constituency tree. T represents the lexical-unit of the target DNI

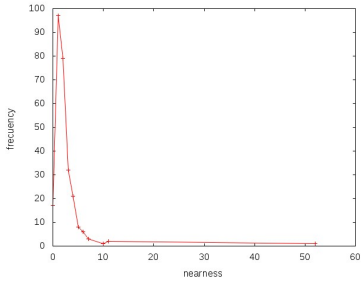

Figure 4: Frequency distribution of the different values of **nearness** in the training data

4.2 Morpho-syntactic and Semantic Agreement

Anaphora and coreference solvers usually apply morpho-syntactic and semantic agreement tests. These constraints check for the consistency between the properties of the target entities and the referents. Several agreement tests such as gender, number or semantic class can be applied. Since most of these tests can not be applied to this task, in this work we have studied part of speech and semantic type agreement.

Semantic Type: To extract the semantic type of the filler of a frame element, we first perform a very simple Word Sense Disambiguation (WSD) process assigning to each word, whenever possible, the most frequent sense of WordNet (Fellbaum, 1998). This heuristic has been used frequently as a baseline in the evaluation of WSD systems. As WordNet senses have been mapped to several ontologies, this disambiguation method allows us to label the documents with ontological features that can work as semantic types. In this work we have used the Top Ontology (TO) (Álvez et al., 2008). We assign to each filler the ontological features of its syntactic head. In this way, we can learn from the training data and for each frame element a probability distribution of its semantic types. Table 2 contains some examples.

Part of Speech: We also calculate the probability distribution of the part of speech (POS) of the head of the fillers similarly as for the semantic types.

Frame#FrameElement	SemanticType	Probability
Expectation#Cognizer	Human	0.93
	Group	0.07
Residence#Location	Building	0.77
	Place	0.33
Attempt#Goal	Purpose	0.41
	UnboundEvent	0.37
	Object	0.13
	Part	0.09

Table 2: Some examples of semantic types assigned to frame elements.

4.3 Discursive

Recency: Recent entities are more likely to be a coreferent than more distant ones. This fact can be easily represented as the sentence distance between the lexical-unit of the target DNI and its referent. This feature has been used frequently not only in coreference and anaphora resolution but also in implicit argument resolution. (Gerber and Chai, 2010) noticed that the vast majority of the fillers of an implicit argument can be found either in the same sentence of the predicate or in the two preceding sentences. In our training data, this fact accounts for 70% of cases. Moreover, only around 2% of the fillers are located in posterior sentences. Figure 5 presents a frequency distribution of the different **recency** values.

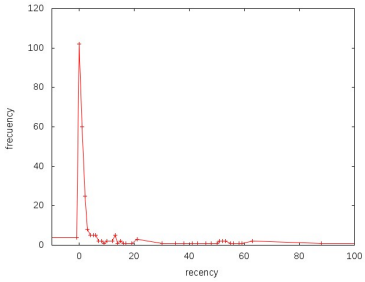

Figure 5: Frequency distribution of the different values of **recency** in the training data

filler LU	dialogue	monologue
dialogue	77.8%	5.4%
monologue	22.2%	94.6%

Table 3: Dialogue vs. monologue distributions

Dialogue: Since the corpus data consists of different chapters of a novel, it contains many dialogues inserting a narrative monologue. The resolution of pronoun and coreference in dialogues dealing with a multi-party discourse have been largely studied (e.g. (Byron and Stent, 1998; Poesio et al., 2006; Stent and Bangalore, 2010)). In this work, we just studied how referents are maintained with respect the two different levels of discourse. Table 3 shows that, in the vast majority of cases, both lexical-unit and filler belong to the same level of discourse[2]. Consequently, this fact can be used to promote those candidates that are at the same discourse level of the lexical-unit of the target DNI.

4.4 Coreference chains

An important source of evidence for anaphora resolution is the focus. The entity or topic which is salient in a particular part of the discourse is the most likely to be coreferred in the same part of the discourse. Thus, given a coreference annotation of a document it is possible to know how the focus varies along the discourse. As we explained in Section 3, the training data contains a full coreference annotation that we use to study three sources of evidence related to both focus and coreference chains.

Non singleton: Using the same training data, (Silberer and Frank, 2012) found that 72% of the DNIs are linked to referents that belong to non-singleton coreference chains. This means that candidate entities that are mentioned just once are less likely to be a referent filler of an implicit argument.

[2]Moreover, as expected, it is more frequent to refer from a monologue to a dialogue entity than the opposite.

Focus: The **focus** refers to the entities that are most likely to be coreferred in a given point in the discourse (Sidner, 1978; Grosz and Sidner, 1986). Now, we study if this is also satisfied for DNI referents by checking if the filler of a DNI corresponds to the **focus** of a near context. We define the **focus** in a near context as follows. Consider the following definitions:

- F is the mention of an entity that is annotated as a filler of a target DNI.
- T is the lexical-unit of the target DNI.
- E is any entity between F and T.
- F-1 is the previous mention of F in the coreference chain.
- Nf is the number of mentions of F from F-1 to T.
- Ne is the number of mentions of E from F-1 to T.

If F-1 is the previous mention of F in the coreference chain, then Nf is equal to two. If there are no previous mentions of F, then F-1 is equal to F, and Nf is equal to one. F is the focus of the near context of T if and only if there is no E complying with Ne>Nf.

From our training data, we also observe that the **focus** matches the filler of a DNI in 72% of the cases.

Centering: **Centering** (Grosz et al., 1995; Brennan et al., 1987) is a theory that tracks the continuity of the focus to explain the coherence of the referential entities in a discourse. The theory establishes three different types of focus transition depending on the relation within the previous focus, Cb(Un-1), the actual focus, Cb(Un), and the element that is most likely to be the focus, Cp(Un), according to its grammatical function. Figure 6 shows the three different kinds of **centering** transitions.

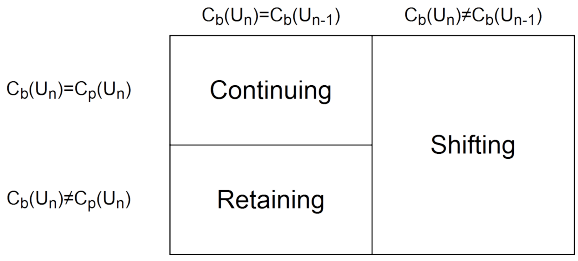

Figure 6: Types of **centering** transitions

The theory establishes that the most common transition is **continuing**. The second most common transition is **retaining** and the least common transition is **shifting**. Applying this schema to the training data, we found that the following probability distribution:

- Continuing: 41.0%
- Retaining: 25.2%
- Shifting: 18.8%
- Other: 15.0%

Since in the DNI filling task the referents can be of any kind of POS and the grammatical function only takes into account nouns or pronouns, the **centering** theory is not always applicable. When the filler is not a noun or a pronoun we have created a fake **centering** category called **other**. Thus, according to the training data, it seems that the preference order of the transitions matches the original theory being **continuing** the most common transition.

5 Experiments

In the previous section we have proposed an adaptation to the implicit argument filling task of some theories traditionally applied to capture evidence for anaphora and coreference resolution. Since the implicit role reference is a special case of coreference, we expect a similar behaviour also for this case. In fact, our analysis using the training data of SemEval seems to confirm our initial hypothesis. In order to evaluate the potential utility of these sources of evidence we have performed a set of experiments using

the SemEval-2010 Task 10 testing-data. In this section, we describe our strategy for solving the implicit arguments and the scorer system used in the evaluation.

Processing Steps: Any system presented to the implicit argument resolution subtask had to follow the following three steps:

1. Select the *frame elements* that are Null Instantiations.
2. Decide if the *null instantiations* are Definite.
3. In case of definite *null instantiation*, locate the corresponding filler.

For the first two steps, we have followed the strategy proposed by (Laparra and Rigau, 2012). This method learns patterns of concurrent *Frame Elements* from explicit annotations. The most common patterns help to identify a missing FE when the rest of the FEs appears explicitly. Following this simple approach, 66% of DNIs in the testing data can be recognized correctly.

For the last step, we have modelled the sources of evidence presented previously as features to train a Naive-Bayes algorithm. We applied a maximum-likelihood method without any smoothing function. Thus, having a set of features *f*, for each DNI we select as filler the candidate *c* that satisfies:

$$\arg\max P(c) \prod_i P(f_i|c)$$

Non-singleton, focus and centering features require a coreference annotation of the document to be analysed. As we explain in Section 3, the training data of the SemEval task contains manually annotated coreference chains that can be used to exploit these features. However, as the testing data does not contain this type of annotations, we applied an automatic coreference resolution system. We used the software provided by Stanford [3]. In the following experiments, we present the results obtained using manual and predicted coreference.

Score measures: The scorer provided for the NI SemEval subtask works slightly different than previous scorers for traditional SRL tasks. Since the participants can appear repeatedly along the document, the scorer needs to take into account the coreference chains of the possible fillers. Thus, if a system selects any of the mentions of the correct filler, the scorer will count it as correct. For this purpose, the dataset provides a full manual coreference annotation. In this subtask, the NI linking precision is defined as the number of all true positive links divided by the number of links made by a system. NI linking recall is defined as the number of true positive links divided by the number of links between an NI and its equivalence set in the gold standard. NI linking F-Score is then calculated as their harmonic mean.

However, since any prediction including the head of the correct filler is scored positively, selecting very large spans of text would obtain very good results[4]. For example, [*madam*] and [*no good will, madam*] would be evaluated as positive results for a [*madam*] gold-standard annotation. Therefore, the scorer also computes the overlap (Dice coefficient) between the words in the predicted filler (P) of an NI and the words in the gold standard one (G):

$$\textbf{NI linking overlap} = \frac{2|P \cap G|}{|P| + |G|}$$

Results on the SemEval-2010 test: Table 4 shows available precision, recall, F-score and overlapping figures of the different systems using predicted and gold-standard coreference chains[5]. Our simple strategy clearly outperforms (Tonelli and Delmonte, 2010) in terms of both precision and recall. (Chen et al., 2010) seems to solve more accurately but a more limited number of cases. We also include the results from (Silberer and Frank, 2012) obtained when using for training a larger corpus extended heuristically (best) and the results obtained with no additional training data (no extra train). Our approach obtains better results in all the cases except when they use extended training data with the gold-standard

[3] `http://nlp.stanford.edu/software/dcoref.shtml`

[4] In particular, returning the whole document would obtain perfect precision and recall.

[5] Surprisingly, previous research do not report results of overlapping. The authors of (Laparra and Rigau, 2012) kindly provided their overlapping results through personal communication.

coreference chains. In this case, our approach seems to achieve a similar performance but without exploiting extra training data. Apparently, (Laparra and Rigau, 2012) presents better results but, as we explained previously, a low overlapping score means vague answers. Although our approach outperforms previous approaches, such a low figures clearly reflect the inherent difficulty of the task.

	Auto Coref				GS Coref			
System	**P**	**R**	**F1**	**Over.**	**P**	**R**	**F1**	**Over.**
(Tonelli and Delmonte, 2010)	-	-	0.01	-				
(Chen et al., 2010)	**0.25**	0.01	0.02	-				
(Tonelli and Delmonte, 2011)	0.13	0.06	0.08	-				
(Silberer and Frank, 2012) no extra train	0.06	0.09	0.07	-	-	-	0.13	-
(Silberer and Frank, 2012) best	0.09	0.11	0.10	-	-	-	**0.18**	-
(Laparra and Rigau, 2012)	0.15	**0.25**	**0.19**	0.54				
This work	0.14	0.18	0.16	**0.89**	0.16	0.20	**0.18**	0.90

Table 4: Results using SemEval-2010 dataset.

DNI linking experiment: In order to check the sources of evidence independently of the rest of processes, we have performed a second experiment where we assume perfect results for the first two steps. In other words, we apply our DNI filling strategy just to the correct DNIs in the document. Table 5 shows the relevance of a correct DNI identification (the first two steps of the process). Once again, without extra training data our strategy outperforms the model of (Silberer and Frank, 2012)[6]. Again, when using extended training data their model seems to perform similar to ours.

	Auto Coref				GS Coref			
System	**P**	**R**	**F1**	**Over.**	**P**	**R**	**F1**	**Over.**
(Silberer and Frank, 2012) no extra train					0.26	**0.25**	0.25	-
(Silberer and Frank, 2012) best					0.31	**0.25**	**0.28**	-
This work	0.30	0.22	0.26	0.89	**0.33**	0.24	**0.28**	0.89

Table 5: Results using SemEval-2010 dataset on the correct DNIs.

Ablation tests: Table 6 presents the results using the gold-standard coreference, when leaving out a type of feature one at a time. The table empirically demonstrates that all feature types contribute positively to solve this task. The morpho-syntactic and semantic agreement seem to be the most relevant evidence in terms of precision and recall. That is, identifying the head of the correct filler. On the other hand, syntactic features are the most relevant to detect the correct span of the fillers.

Source Set	**P**	**R**	**F1**	**Over.**
all	0.33	0.24	0.28	0.89
no-coref	0.30	0.22	0.25	0.86
no-semagree	0.22	0.22	0.22	0.90
no-discursive	0.29	0.22	0.25	0.82
no-syntactic	0.28	0.21	0.24	0.75

Table 6: Ablation tests using the gold-standard coreference.

6 Discussion

In order to analyse the limits of the different types of evidence, we used as reference the results obtained using the gold-standard DNIs and coreference chains (see table 5). As an overall remark, all previous works facing this task agree on the sparsity of the training data. We also observed that this problem affects all sources of evidence we have studied, especially the agreement of semantic types.

[6]The rest of the systems do not perform any experiments with gold-standard DNI identification.

Data sparsity for semantic types: The semantic types do not cover the full set of frame elements. The testing data contains a total of 209 different Frame#FrameElements. 73 of them (out of 35%) do not appear on the training data. Another problem appears when the FEs have too many different semantic types with very similar probabilities. Without enough information to discriminate correctly the filler, this source of evidence becomes damaging (see table 7).

Outside the same sentence: Recency strongly prioritises the window formed by the same sentence of the lexical-unit of the target DNI and the two previous sentences. However, in 19% of the cases the filler belongs to a sentence outside that window. Furthermore, syntactic based evidences rely on relations between entities in the same sentence. Obviously, adding an artificial node covering the whole document analysis is quite arbitrary. Table 8 shows how the performance of our approach decreases strongly when the filler and the lexical-unit are in different sentences.

P	R	F1	Over.
0.21	0.09	0.13	0.61

Table 7: Performance of FE having more than 5 semantic types

same sentence				another sentence			
P	**R**	**F1**	**Over.**	**P**	**R**	**F1**	**Over.**
0.50	0.34	0.40	0.87	0.20	0.16	0.18	0.96

Table 8: Performance when the filler and the lexical-unit are in the same sentence or in another one

Discursive structure: The particular structure of the documents can also affect seriously the performance of the sources of evidence. Table 9 presents the results on contexts with at least 10% of entities on monologue or dialogue. According to the recency feature, each context is formed by the sentence of the lexical-unit of the target DNI and the two previous sentences. We can observe that the results on mixed contexts are better than in general. Obviously, dialogue features are totally useless in contexts with only monologues or only dialogues.

Singleton fillers: Most of the fillers are entities that belong to a coreference chain. Therefore, these cases heavily depends on a correct coreference annotation. This is why worse results are obtained when using predicted coreferent chains. Table 10 shows the results when the filler belongs or not to a coreference chain. It is important to remind that in this work we have adapted a set of sources of evidence and theories traditionally used is anaphora and coreference resolution. Originally these theories focused just on noun and pronoun entities.

P	R	F1	Over.
0.38	0.29	0.33	0.93

Table 9: Performance in mixed contexts with at least 10% of entities of each level

coref-chain				no-coref-chain			
P	**R**	**F1**	**Over.**	**P**	**R**	**F1**	**Over.**
0.45	0.35	0.39	0.94	0.06	0.04	0.05	0.31

Table 10: Performance when the filler belongs to a coreference-chain or not

7 Conclusions and Future Work

We have presented a first attempt to study the behaviour of traditional coreference and anaphora models for the implicit argument resolution task, a special case of coreference. Our analysis shows that these theories and models can be successfully applied as sources of evidence in an existing dataset. In fact, their joint combination improves state of the art results.

However, the sources of evidence presented in this work are adaptations that focus on nominal entities and pronouns, and on relations within entities and referents belonging to the same sentence. It seems that for these cases it is possible to capture useful evidence. But for the rest (singletons, non nominal POS, beyond sentence boundaries, etc.), further investigation is needed.

We have also observed, that the training data of the SemEval-2010 task 10 is too small. Possibly, the results could be improved using an extended training data (Silberer and Frank, 2012).

Following the line of research presented by Roth and Frank (Roth and Frank, 2012b,a) we will study the influence between the implicit arguments resolution and the predicate alignment.

Finally, we plan to perform a similar study on the NomBank dataset (Gerber and Chai, 2010).

8 Acknowledgment

We are grateful to Rodrigo Agerri and the anonymous reviewers for their insightul comments. This work has been partially funded by SKaTer (TIN2012-38584-C06-02), OpeNER (FP7-ICT-2011-SME-DCL-296451) and NewsReader (FP7-ICT-2011-8-316404).

References

Álvez, J., J. Atserias, J. Carrera, S. Climent, E. Laparra, A. Oliver, and G. Rigau (2008). Complete and consistent annotation of wordnet using the top concept ontology. In *LREC*.

Baker, C. F., C. J. Fillmore, and J. B. Lowe (1998). The berkeley framenet project. In *COLING-ACL*, pp. 86–90.

Brennan, S. E., M. W. Friedman, and C. J. Pollard (1987). A centering approach to pronouns. In *Meeting of the Association for Computational Linguistics*.

Burger, J. D. and D. Connolly (1992). Probabilistic resolution of anaphoric reference. In *AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, Cambridge, MA, USA, pp. 17–24.

Byron, D. K. and A. Stent (1998). A Preliminary Model of Centering in Dialog. In *Meeting of the Association for Computational Linguistics*, pp. 1475–1477.

Chen, D., N. Schneider, D. Das, and N. A. Smith (2010). Semafor: Frame argument resolution with log-linear models. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, Stroudsburg, PA, USA, pp. 264–267. Association for Computational Linguistics.

Fellbaum, C. (Ed.) (1998). *WordNet: an electronic lexical database*. MIT Press.

Gerber, M. and J. Y. 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*, ACL '10, Stroudsburg, PA, USA, pp. 1583–1592. Association for Computational Linguistics.

Gildea, D. and D. Jurafsky (2000). Automatic labeling of semantic roles. In *ACL*.

Grosz, B. J., A. K. Joshi, and S. Weinstein (1995). Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics 21*, 203–225.

Grosz, B. J. and C. L. Sidner (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics 12*, 175–204.

Hobbs, J. R. (1977). Pronoun resolution. *Intelligence/sigart Bulletin*, 28–28.

Laparra, E. and G. Rigau (2012). Exploiting explicit annotations and semantic types for implicit argument resolution. In *6th IEEE International Conference on Semantic Computing*, ICSC '12, Palermo, Italy.

Meyers, A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman (2004, May 2 - May 7). The nombank project: An interim report. In A. Meyers (Ed.), *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, Boston, Massachusetts, USA, pp. 24–31. Association for Computational Linguistics.

Palmer, M., D. A. Dahl, R. J. Schiffman, L. Hirschman, M. C. Linebarger, and J. Dowding (1986). Recovering implicit information. In *ACL*, pp. 10–19.

Poesio, M., A. Patel, and B. D. Eugenio (2006). Discourse Structure and Anaphora in Tutorial Dialogues: An Empirical Analysis of Two Theories of the Global Focus. *Research on Language and Computation 4*, 229–257.

Reinhart, T. (1976). *The syntactic domain of anaphora*. MIT Linguistics Dissertations. Massachusetts Institute of Technology.

Roth, M. and A. Frank (2012a). Aligning predicate argument structures in monolingual comparable texts: A new corpus for a new task. In *Proceedings of *SEM 2012: The First Conference on Lexical and Computational Semantics*, Montreal, Canada.

Roth, M. and A. Frank (2012b). Aligning predicates across monolingual comparable texts using graph-based clustering. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processign (EMNLP)*, Jeju, Republic of Korea.

Ruppenhofer, J., P. Gorinski, and C. Sporleder (2011). In search of missing arguments: A linguistic approach. In G. Angelova, K. Bontcheva, R. Mitkov, and N. Nicolov (Eds.), *Recent Advances in Natural Language Processing, RANLP 2011, 12-14 September, 2011, Hissar, Bulgaria*, pp. 331–338. RANLP 2011 Organising Committee.

Ruppenhofer, J., C. Sporleder, R. Morante, C. Baker, and M. Palmer (2009). Semeval-2010 task 10: linking events and their participants in discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, DEW '09, Stroudsburg, PA, USA, pp. 106–111. Association for Computational Linguistics.

Sidner, C. L. (1978). The use of focus as a tool for disambiguation of definite noun phrases. In *Proceedings of the 1978 workshop on Theoretical issues in natural language processing*, TINLAP '78, Stroudsburg, PA, USA, pp. 86–95. Association for Computational Linguistics.

Silberer, C. and A. Frank (2012). Casting implicit role linking as an anaphora resolution task. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, Montréal, Canada, pp. 1–10. Association for Computational Linguistics.

Stent, A. J. and S. Bangalore (2010). Interaction between dialog structure and coreference resolution. In *IEEE Workshop on Spoken Language Technology*.

Tetreault, J. R. (2002). Implicit role reference. In *International Symposium on Reference Resolution for Natural Language Processing*, Alicante, Spain, pp. 109–115.

Tonelli, S. and R. 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*, SemEval '10, Stroudsburg, PA, USA, pp. 296–299. Association for Computational Linguistics.

Tonelli, S. and R. Delmonte (2011, June). Desperately seeking implicit arguments in text. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, Portland, Oregon, USA, pp. 54–62. Association for Computational Linguistics.

Whittemore, G., 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*, ACL '91, Stroudsburg, PA, USA, pp. 17–24. Association for Computational Linguistics.