SemEval-2010 Task 10: Linking Events and Their Participants in Discourse

Josef Ruppenhofer and **Caroline Sporleder**
Computational Linguistics
Saarland University
{josefr,csporled}@coli.uni-sb.de

Roser Morante
CNTS
University of Antwerp
Roser.Morante@ua.ac.be

Collin Baker
ICSI
Berkeley, CA 94704
collin@icsi.berkeley.edu

Martha Palmer
Department of Linguistics
University of Colorado at Boulder
martha.palmer@colorado.edu

Abstract

In this paper, we describe the SemEval-2010 shared task on "Linking Events and Their Participants in Discourse". This task is a variant of the classical semantic role labelling task. The novel aspect is that we focus on linking local semantic argument structures across sentence boundaries. Specifically, the task aims at linking locally uninstantiated roles to their co-referents in the wider discourse context (if such co-referents exist). This task is potentially beneficial for a number of NLP applications and we hope that it will not only attract researchers from the semantic role labelling community but also from co-reference resolution and information extraction.

1 Introduction

Semantic role labelling (SRL) has been defined as a sentence-level natural-language processing task in which semantic roles are assigned to the syntactic arguments of a predicate (Gildea and Jurafsky, 2002). Semantic roles describe the function of the participants in an event. Identifying the semantic roles of the predicates in a text allows knowing who did what to whom when where how, etc.

SRL has attracted much attention in recent years, as witnessed by several shared tasks in Senseval/SemEval (Màrquez et al., 2007; Litkowski, 2004; Baker et al., 2007; Diab et al., 2007), and CoNLL (Carreras and Màrquez, 2004; Carreras and Màrquez, 2005; Surdeanu et al., 2008). The state-of-the-art in semantic role labelling has now advanced so much that a number of studies have shown that automatically inferred semantic argument structures can lead to tangible performance gains in NLP applications such as information extraction (Surdeanu et al., 2003), question answering (Shen and Lapata, 2007) or recognising textual entailment (Burchardt and Frank, 2006).

However, semantic role labelling as it is currently defined also misses a lot of information that would be beneficial for NLP applications that deal with text understanding (in the broadest sense), such as information extraction, summarisation, or question answering. The reason for this is that SRL has traditionally been viewed as a sentence-internal task. Hence, relations between different local semantic argument structures are disregarded and this leads to a loss of important semantic information.

This view of SRL as a sentence-internal task is partly due to the fact that large-scale manual annotation projects such as FrameNet[1] and PropBank[2] typically present their annotations lexicographically by lemma rather than by source text. Furthermore, in the case of FrameNet, the annotation effort did not start out with the goal of exhaustive corpus annotation but instead focused on isolated instances of the target words sampled from a very large corpus, which did not allow for a view of the data as 'full-text annotation'.

It is clear that there is an interplay between local argument structure and the surrounding discourse (Fillmore, 1977). In early work, Palmer et al. (1986) discussed filling null complements from context by using knowledge about individual predicates and ten-

[1] http://framenet.icsi.berkeley.edu/

[2] http://verbs.colorado.edu/~mpalmer/projects/ace.html

Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, pages 106–111,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

dencies of referential chaining across sentences. But so far there have been few attempts to find links between argument structures across clause and sentence boundaries explicitly on the basis of semantic relations between the predicates involved. Two notable exceptions are Fillmore and Baker (2001) and Burchardt et al. (2005). Fillmore and Baker (2001) analyse a short newspaper article and discuss how frame semantics could benefit discourse processing but without making concrete suggestions of how to model this. Burchardt et al. (2005) provide a detailed analysis of the links between the local semantic argument structures in a short text; however their system is not fully implemented either.

In the shared task, we intend to make a first step towards taking SRL beyond the domain of individual sentences by linking local semantic argument structures to the wider discourse context. In particular, we address the problem of finding fillers for roles which are neither instantiated as direct dependents of our target predicates nor displaced through long-distance dependency or coinstantatiation constructions. Often a referent for an uninstantiated role can be found in the wider context, i.e. in preceding or following sentences. An example is given in (1), where the CHARGES role (ARG2 in PropBank) of *cleared* is left empty but can be linked to *murder* in the previous sentence.

(1) In a lengthy court case the defendant was tried for murder. In the end, he was cleared.

Another very rich example is provided by (2), where, for instance, the experiencer and the object of jealousy are not overtly expressed as syntactic dependents of the noun *jealousy* but can be inferred to be Watson and the speaker, Holmes, respectively.

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

NIs are also very frequent in clinical reports. For example, in (3) the EXPERIENCER role of "cough", "tachypnea", and "breathing" can be linked to "twenty-two month old". Text mining systems in the biomedical domain focus on extracting relations between biomedical entities and information about patients. It is important that these systems extract information as accurately as possible. Thus, finding co-referents for NIs is also very relevant for improving results on mining relations in biomedical texts.

(3) Twenty-two month old with history of recurrent right middle lobe infiltrate. Increased cough, tachypnea, and work of breathing.

In the following sections we describe the task in more detail. We start by providing some background on null instantiations (Section 2). Section 3 gives an overview of the task, followed by a description of how we intend to create the data (Section 4). Section 5 provides a short description of how null instantiations could be resolved automatically given the provided data. Finally, Section 6 discusses the evaluation measures and we wrap up in Section 7.

2 Background on Null Instantiation

The theory of null complementation used here is the one adopted by FrameNet, which derives from the work of Fillmore (1986).[3] Briefly, omissions of core arguments of predicates are categorised along two dimensions, the licensor and the interpretation they receive. The idea of a licensor refers to the fact that either a particular lexical item or a particular grammatical construction must be present for the omission of a frame element (FE) to occur. For instance, the omission of the agent in (4) is licensed by the passive construction.

(4) No doubt, mistakes were made $0^{Protagonist}$.

The omission is a constructional omission because it can apply to any predicate with an appropriate semantics that allows it to combine with the passive construction. On the other hand, the omission in (5) is lexically specific: the verb *arrive* allows the Goal to be unspecified but the verb *reach*, also a member of the Arriving frame, does not.

(5) We arrived 0^{Goal} at 8pm.

The above two examples also illustrate the second major dimension of variation. Whereas, in (4) the protagonist making the mistake is only existentially bound within the discourse (instance of indefinite null

[3] Palmer et al.'s (1986) treatment of uninstantiated 'essential roles' is very similar (see also Palmer (1990)).

instantiation, INI), the Goal location in (5) is an entity that must be accessible to speaker and hearer from the discourse or its context (definite null instantiation, DNI). Finally note that the licensing construction or lexical item fully and reliably determines the interpretation. Missing by-phrases always have an indefinite interpretation and whenever *arrive* omits the Goal lexically, the Goal has to be interpreted as definite, as it is in (5).

The import of this classification to the task here is that we will concentrate on cases of DNI whether they are licensed lexically or constructionally.

3 Task Description

We plan to run the task in the following two modes:

Full Task For the full task we supply a test set in which the target words are marked and labelled with the correct sense (i.e. frame).[4] The participants then have to:

1. find the overt semantic arguments of the target (role recognition)
2. label them with the correct role (role labelling)
3. recognize definite null instantiations and find links to antecedents in the wider context (NI linking)

NIs only In the second mode, participants will be supplied with a test set which is annotated with gold standard local semantic argument structure.[5] The task is then restricted to recognizing that a core role is missing, ascertaining that it must have a definite interpretation and finding a filler for it (i.e., sub-task 3 from the full task).

The full task and the null instantiation linking task will be evaluated separately. By setting up a SRL task, we expect to attract participants from the established SRL community. Furthermore, by allowing participants to only address the second task, we hope to also attract researchers from areas such as coreference resolution or information extraction who do not want to implement a complete SRL system. We also plan to provide the data with both FrameNet and PropBank style annotations to encourage researchers from both areas to take part.

[4] We supply the correct sense to ensure that all systems use the same role inventory for each target (i.e., the role inventory associated with the gold standard sense). This makes it easier to evaluate the systems consistently with respect to role assignments and null instantiation linking, which is our main focus.

[5] The training set is identical for both set-ups and will contain the full annotation, i.e., frames, semantic roles and their fillers, and referents of null instantiations in the wider context (see Section 4 for details).

4 Data

The data will come from one of Arthur Conan Doyle's fiction works. We chose fiction rather than news because we believe that fiction texts with a linear narrative generally contain more context-resolvable null instantiations. They also tend to be longer and have a simpler structure than news texts which typically revisit the same facts repeatedly at different levels of detail (in the so-called 'inverted pyramid' structure) and which mix event reports with commentary and evaluation, thus sequencing material that is understood as running in parallel. Fiction texts should lend themselves more readily to a first attempt at integrating discourse structure into semantic role labeling. We chose Conan Doyle's work because most of his books are not subject to copyright restrictions anymore, which allows us to freely release the annotated data.

We plan to make the data sets available with both FrameNet and PropBank semantic argument annotation, so that participants can choose which framework they want to work in. The annotations will originally be made using FrameNet-style and will later be mapped semi-automatically to PropBank annotations. The data set for the FrameNet version of the task will be built at Saarland University, in close co-operation with the FrameNet team in Berkeley. We aim for the same density of annotation as is exhibited by FrameNet's existing full-text annotation[6] and are currently investigating whether the semantic argument annotation can be done semi-automatically, e.g., by starting the annotation with a run of the Shalmaneser role labeller (Erk and Padó, 2006), whose output is then corrected and expanded manually. To ensure a high annotation quality, at least part of the data will be annotated by two annotators and then manually adjudicated. We also provide detailed annotation guidelines (largely following the FrameNet

[6] `http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=84`

guidelines) and any open questions are discussed in a weekly annotation meeting.

For the annotation of null instantiations and their links to the surrounding discourse we have to create new guidelines as this is a novel annotation task. We will adopt ideas from the annotation of co-reference information, linking locally unrealised roles to all mentions of the referents in the surrounding discourse, where available. We will mark only identity relations but not part-whole or bridging relations between referents. The set of unrealised roles under consideration includes only the core arguments but not adjuncts (peripheral or extra-thematic roles in FrameNet's terminology). Possible antecedents are not restricted to noun phrases but include all constituents that can be (local) role fillers for some predicate plus complete sentences (which can sometimes fill roles such as MESSAGE).

The data-set for PropBank will be created by mapping the FrameNet annotations onto PropBank and NomBank labels. For verbal targets, we use the Semlink[7] mappings. For nominal targets, there is no existing hand-checked mapping between FrameNet and NomBank but we will explore a way of building a FrameNet - NomBank mapping at least for eventive nouns indirectly with the help of Semlink. This would take advantage of the fact that PropBank verbs and eventive NomBank nouns both have a mapping to VerbNet classes, which are referenced also by Semlink. Time permitting, non-eventive nouns could be mapped manually. For FrameNet targets of other parts of speech, in particular adjectives and prepositions, no equivalent PropBank-style counterparts will be available. The result of the automatic mappings will be partly hand-checked. The annotations resolving null instantiations need no adjustment.

We intend to annotate at least two data sets of around 4,000 words. One set for testing and one for training. Because we realise that the training set will not be large enough to train a semantic role labelling system on it, we permit the participants to boost the training data for the SRL task by making use of the existing FrameNet and PropBank corpora.[8]

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

[8] This may require some genre adaption but we believe this is feasible.

5 Resolving Null Instantiations

We conceive of null instantiation resolution as a three step problem. First, one needs to determine whether a core role is missing. This involves looking up which core roles are overtly expressed and which are not.

In the second step, one needs to determine what licenses an omission and what its interpretation is. To do this, one can use rules and heuristics based on various syntactic and lexical facts of English. As an example of a relevant syntactic fact, consider that subjects in English can only be omitted when licensed by a construction. One such construction is the imperative (e.g. *Please, sit down*). Since this construction also specifies that the missing referent must be the addressee of the speaker of the imperative, it is clear what referent one has to try to find.

As for using lexical knowledge, consider omissions of the Goods FE of the verb *steal* in the Theft frame. FrameNet annotation shows that whenever the Goods FE of *steal* is missing it is interpreted indefinitely, suggesting that a new instance of the FE being missing should have the same interpretation.

More evidence to the same effect can be derived using Ruppenhofer's (2004) observation that the interpretation of a lexically licensed omission is definite if the overt instances of the FE have mostly definite form (i.e. have definite determiners such as *that*, *the* , *this*), and indefinite if they are mostly indefinite (i.e. have bare or indefinite determiners such as *a(n)* or *some*). The morphology of overt instances of an FE could be inspected in the FrameNet data, or if the predicate has only one sense or a very dominant one, then the frequencies could even be estimated from unannotated corpora.

The third step is linking definite omissions to referents in the context. This linking problem could be modelled as a co-reference resolution task. While the work of Palmer et al. (1986) relied on special lexicons, one might instead want to learn information about the semantic content of different role fillers and then assess for each of the potential referents in the discourse context whether their semantic content is close enough to the expected content of the null instantiated role.

Information about the likely fillers of a role can be obtained from annotated data sets (e.g., FrameNet or PropBank). For instance, typical fillers of the

CHARGES role of *clear* might be *murder*, *accusations*, *allegations*, *fraud* etc. The semantic content of the role could then be represented in a vector space model, using additional unannotated data to build meaning vectors for the attested role fillers. Meaning vectors for potential role fillers in the context of the null instantiation could be built in a similar fashion. The likelihood of a potential filler filling the target role can then be modelled as the distance between the meaning vector of the filler and the role in the vector space model (see Padó et al. (2008) for a similar approach for semi-automatic SRL).

We envisage that the manually annotated null instantiated data can be used to learn additionally heuristics for the filler resolution task, such as information about the average distance between a null instantiation and its most recent co-referent.

6 Evaluation

As mentioned above we allow participants to address either the full role recognition and labelling task plus the linking of null instantiations or to make use of the gold standard semantic argument structure and look only at the null instantiations. We also permit systems to perform either FrameNet or PropBank style SRL. Hence, systems can be entered for four subtasks which will be evaluated separately:

- full task, FrameNet
- null instantiations, FrameNet
- full task, PropBank
- null instantiations, PropBank

The focus for the proposed task is on the null instantiation linking, however, for completeness, we also evaluate the standard SRL task. For role recognition and labelling we use a standard evaluation set-up, i.e., for role recognition we will evaluate the accuracy with respect to the manually created gold standard, for role labelling we will evaluate precision, recall, and F-Score.

The null instantiation linkings are evaluated slightly differently. In the gold standard, we will identify referents for null instantiations in the discourse context. In some cases, more than one referent might be appropriate, e.g., because the omitted argument refers to an entity that is mentioned multiple times in the context. In this case, a system should be given credit if the null instantiation is linked to any of these expressions. To achieve this we create equivalence sets for the referents of null instantiations. If the null instantiation is linked to any item in the equivalence set, the link is counted as a true positive. We can then define **NI linking precision** as the number of all true positive links divided by the number of links made by a system, and **NI linking recall** as the number of true positive links divided by the number of links between a null instantiation and its equivalence set in the gold standard. **NI linking F-Score** is then the harmonic mean between NI linking precision and recall.

Since it may sometimes be difficult to determine the correct extend of the filler of an NI, we score an automatic annotation as correct if it includes the head of the gold standard filler in the predicted filler. However, in order to not favour systems which link NIs to excessively large spans of text to maximise the likelihood of linking to a correct referent, we introduce a second evaluation measure, which computes the overlap (Dice coefficient) between the words in the predicted filler (P) of a null instantiation and the words in the gold standard one (G):

$$\textbf{NI linking overlap} = \frac{2|P \cap G|}{|P| + |G|} \qquad (6)$$

Example (7) illustrates this point. The verb *won* in the second sentence evokes the Finish_competition frame whose COMPETITION role is null instantiated. From the context it is clear that the competition role is semantically filled by *their first TV debate* (head: *debate*) and *last night's debate* (head: *debate*) in the previous sentences. These two expressions make up the equivalence set for the COMPETITION role in the last sentence. Any system that would predict a linkage to a filler that covers the head of either of these two expressions would score a true positive for this NI. However, a system that linked to *last night's debate* would have an NI linking overlap of 1 (i.e., 2*3/(3+3)) while a system linking the whole second sentence *Last night's debate was eagerly anticipated* to the NI would have an NI linking overlap of 0.67 (i.e., 2*3/(6+3))

(7) US presidential rivals Republican John McCain and Democrat Barack Obama have yesterday evening attacked each other over

foreign policy and the economy, in [their first TV debate]$_{Competition}$. [Last night's debate]$_{Competition}$ was eagerly anticipated. Two national flash polls suggest that [Obama]$_{Competitor}$ won$_{Finish_competition}$ $0^{Competition}$.

7 Conclusion

In this paper, we described the SemEval-2010 shared task on "Linking Events and Their Participants in Discourse". With this task, we intend to take a first step towards viewing semantic role labelling not as a sentence internal problem but as a task which should really take the discourse context into account. Specifically, we focus on finding referents for roles which are null instantiated in the local context. This is potentially useful for various NLP applications. We believe that the task is timely and interesting for a number of researchers not only from the semantic role labelling community but also from fields such as co-reference resolution or information extraction.

While our task focuses specifically on finding links between null instantiated roles and the discourse context, we hope that in setting it up, we can stimulate research on the interaction between discourse structure and semantic argument structure in general. Possible future editions of the task could then focus on additional connections between local semantic argument structures (e.g., linking argument structures that refer to the same event).

8 Acknowledgements

Josef Ruppenhofer and Caroline Sporleder are supported by the German Research Foundation DFG (under grant PI 154/9-3 and the Cluster of Excellence Multimodal Computing and Interaction (MMCI), respectively). Roser Morante's research is funded by the GOA project BIOGRAPH of the University of Antwerp.

References

C. Baker, M. Ellsworth, and K. Erk. 2007. SemEval-2007 Task 19: Frame semantic structure extraction. In *Proceedings of SemEval-07*.

A. Burchardt and A. Frank. 2006. Approximating textual entailment with LFG and framenet frames. In *Proceedings of the Second Recognising Textual Entailment Workshop*.

A. Burchardt, A. Frank, and M. Pinkal. 2005. Building text meaning representations from contextually related frames – A case study. In *Proceedings of IWCS-6*.

X. Carreras and Ll. Màrquez. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of CoNLL-04*, pages 89–97.

X. Carreras and Ll. Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of CoNLL-05*, pages 152–164.

M. Diab, M. Alkhalifa, S. ElKateb, C. Fellbaum, A. Mansouri, and M. Palmer. 2007. SemEval-2007 Task 18: Arabic semantic labeling. In *Proc. of SemEval-07*.

K. Erk and S. Padó. 2006. Shalmaneser - a flexible toolbox for semantic role assignment. In *Proceedings of LREC-06*.

C.J. Fillmore and C.F. Baker. 2001. Frame semantics for text understanding. In *Proc. of the NAACL-01 Workshop on WordNet and Other Lexical Resources*.

C.J. Fillmore. 1977. Scenes-and-frames semantics, linguistic structures processing. In Antonio Zampolli, editor, *Fundamental Studies in Computer Science, No. 59*, pages 55–88. North Holland Publishing.

C.J. Fillmore. 1986. Pragmatically controlled zero anaphora. In *Proceedings of the Twelfth Annual Meeting of the Berkeley Liguistics Society*.

D. Gildea and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

K. Litkowski. 2004. SENSEVAL-3 Task: Automatic labeling of semantic roles. In *Proc. of SENSEVAL-3*.

L. Màrquez, L. Villarejo, M. A. Martì, and M. Taulè. 2007. SemEval-2007 Task 09: Multilevel semantic annotation of Catalan and Spanish. In *Proceedings of SemEval-07*.

S. Padó, M. Pennacchiotti, and C. Sporleder. 2008. Semantic role assignment for event nominalisations by leveraging verbal data. In *Proceedings of Coling-2008*.

M. Palmer, D. Dahl, R. Passonneau, L. Hirschman, M. Linebarger, and J. Dowding. 1986. Recovering implicit information. In *Proceedings of ACL-1986*.

M. Palmer. 1990. *Semantic Processing for Finite Domains*. CUP, Cambridge, England.

J. Ruppenhofer. 2004. *The interaction of valence and information structure*. Ph.d., University of California, Berkeley, CA.

D. Shen and M. Lapata. 2007. Using semantic roles to improve question answering. In *Proc. of EMNLP-07*.

M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth. 2003. Using predicate arguments structures for information extraction. In *Proceedings of ACL-2003*.

M. Surdeanu, R. Johansson, A. Meyers, Ll. Màrquez, and J. Nivre. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of CoNLL-2008*, pages 159–177.