Semantic Argument Structure in DiscoursE: The SEASIDE Project

Caroline Sporleder
Computational Linguistics / Cluster of Excellence
Saarland University
66041 Saarbrücken, Germany
csporled@coli.uni-saarland.de

The recently started SEASIDE project is funded for five years (2008-2013) by the German Excellence Initiative as part of Saarland University's Cluster of Excellence on "Multimodal Computing and Interaction". In the project, we aim to bring together two active research areas which both deal with "computing meaning" but currently stand more or less independently next to each other: *discourse processing* and computation of *semantic argument structure*. We expect that both areas will benefit from this: semantic argument information will allow for a more sophisticated representation of discourse meaning, while discourse information can also be beneficial for systems which compute semantic argument structure (i.e. *semantic role labellers*). Eventually we aim for an incremental model of text meaning which can be computed in a robust, data-driven way by utilising and combining information from several levels of linguistic analysis. The model should be sophisticated enough to aid applications such as text mining, information extraction, question answering, and text summarisation.

Discourse processing deals with modelling the meaning of multi-sentence units. Early approaches (e.g. Hobbs et al., 1993) were heavily knowledge-based and, while these systems worked well on small, well-defined domains, they generally did not scale up very well. More recent research largely abandoned the knowledge-based approach in favour of much shallower systems, either rule-based (Polanyi et al., 2004) or machine-learned (Soricut and Marcu, 2003). These systems rely largely on surface cues. While shallow models can be quite successful, they also have clear limitations. For example, progress on discourse parsing has stagnated in the last years and text summarisation is still a challenge, especially from multiple input documents.

Proceedings of the 8th International Conference on Computational Semantics, pages 333–337,
Tilburg, January 2009.

Theories of **semantic argument structure**, such as Frame Semantics, model relations within individual sentences, namely the relation between a lexical item and its semantic arguments such as *agent* or *patient*. During the last five to ten years there has been much research in this area, as witnessed by several large scale projects aimed at providing lexicons and annotated corpora (e.g., FrameNet,[1] PropBank,[2] and SALSA[3]), and numerous shared tasks on semantic role labelling (Baker et al., 2007; Carreras and Màrquez, 2005; Carreras and Màrques, 2004). While the performance of semantic parsers is still lower than that of syntactic parsers, it is now good enough that NLP tasks such as information extraction or question answering can be shown to benefit from automatically computed semantic argument structures (Moschitti et al., 2003; Shen and Lapata, 2007).

While Frame Semantics was originally seen as being grounded in discourse (Fillmore, 1977), its computational treatment has largely been restricted to the sentence level, which may also be due to the fact that annotated data typically consists of sets of individual sentences rather than of running text, though there has been some effort recently to create full text annotations as well. Few studies tried to connect frame semantic annotations across sentences. 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. Burchardt et al. (2005) provide a more detailed analysis of a short text but their system is not fully implemented.

In the SEASIDE project we aim to bridge the gap between discourse processing and semantic argument structure information by (i) enriching semantic role labelling with discourse information, and (ii) enriching discourse models with information about the semantic argument structure of the individual clauses.

Discourse information could be useful for semantic role labelling in a number of ways:

- by integrating discourse features in the models, e.g. information about the rhetorical relations that hold between adjacent sentences, such as *contrast* or *elaboration*, or about the focus structure
- developing statistical models of the roles that are likely to be realised in different contexts

[1] `http://framenet.icsi.berkeley.edu`
[2] `http://verbs.colorado.edu/~mpalmer/projects/ace.html`
[3] `http://www.coli.uni-saarland.de/projects/salsa/`

- utilising frame-to-frame relations as specified by FrameNet to improve frame disambiguation and role labelling
- equipping semantic role labellers with a "memory" to allow them to build text meaning representations incrementally rather than having to start "from scratch" for each target predicate

We also believe that semantic role labelling should not stop at the sentence level; semantic argument structures are often incomplete and linking them across clause boundaries will benefit many NLP tasks. For instance, consider the verb *clear* in example (1). This verb evokes the VERDICT frame which has a role for CHARGES that is not filled locally (i.e., by any of the syntactic constituents in the second sentence) but can be inferred from the preceding sentence, which specifies the charges as *for murder*. Semantic role labelling systems which operate on the sentence-level miss this crucial fact and will be unable to fill the CHARGES role of VERDICT, even though it is present in the discourse context. Systems which can link local semantic argument structures can create more complete meaning representations of a text than semantic role labellers restricted to the local domain. In order to stimulate research in this direction, we are organising a Shared Task at SemEval-2010 on finding links between locally uninstantiated roles and the discourse context.[4] To our knowledge, the data we are creating for this task will be the first publicly available reference data set containing information about global linking of semantic argument structures.

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

While discourse information can be beneficial for the computation of sematic argument structures, the reverse is also true: the semantic argument structures in a text and their relations can provide vital cues about the coherence of the discourse. Incorporating (automatically computed) argument structure information leads to more sophisticated models of discourse structure. Such models encode deeper linguistic information than models based on surface cues, while still being computable in a data-driven fashion. Utilising frame semantic information can, for example, explain why example (23) is perceived as more coherent than (24): The verb *try* evokes the TRY_DEFENDANT frame which is closely linked to the SENTENCING frame evoked by *sentence*. Furthermore the DEFENDANT role of the first frame is

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

co-referent with the CONVICT role of the second frame, and the unrealised COURT role of the SENTENCING frame can be linked to *The High Court* in the following sentence. The discourse in (24), on the other hand, is perceived as less coherent. One reason for this is that there are fewer links between the semantic argument structures in the two sentences. For instance, the LOSE_POSSESSION frame evoked by *lost* cannot be linked easily to TRY_DEFENDANT. Nor are any roles shared between the frames in the two sentences, with the exception of the co-reference between the CONVICT role of SENTENCING (*Dan Talor*) and the DONOR role of LOSE_POSSESSION (*He*). While the absence of obvious semantic argument structure links does not necessarily mean that a text is not coherent,[5] their presence is likely to be a fairly reliable cue for coherence.

(2) (3) Dan Taylor was tried for murder. The High Court sentenced him to life imprisonment.

(4) Dan Taylor was tried for murder. He had lost his car keys.

In the SEASIDE project, we aim to develop models which can compute interconnected semantic argument representations for a given text, enabling us to predict such differences in coherence.

References

Collin Baker, Michael Ellsworth, and Katrin Erk. Semeval-2007 task 19: Frame semantic structure extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104, 2007.

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

Xavier Carreras and Lluís Màrques. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-04)*, pages 89–97, 2004.

Xavier Carreras and Lluís Màrquez. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, Michigan, 2005.

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

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

[5] The sentence sequence *Dan Taylor was tried for murder. He had lost his patience.* for example, might be interpreted as coherent.

Jerry R. Hobbs, Mark Stickel, Douglas Appelt, and Paul Martin. Interpretation as abduction. *Artificial Intelligence*, 63(1-2):69–142, 1993.

Alessandro Moschitti, Paul Morarescu, and Sanda Harabagiu. Open-domain information extraction via automatic semantic labeling. In *Proceedings of FLAIRS*, pages 397–401, 2003.

Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. A rule based approach to discourse parsing. In *Proc. of the 5th SIGDIAL Workshop in Discourse and Dialogue*, pages 108–117, 2004.

Dan Shen and Mirella Lapata. Using semantic roles to improve question answering. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP-07)*, 2007.

Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proc. HLT-NAACL-03*, 2003.