The SALSA Annotation Tool

Katrin Erk and Andrea Kowalski and Sebastian Padó
Department of Computational Linguistics
Saarland University
66123 Saarbrücken
erk, kowalski, pado@coli.uni-sb.de

Abstract

The SALSA annotation tool supports the graphical annotation of a treebank with semantic roles in the frame semantics paradigm. The tool, which takes corpora in the TIGER XML format as input, supports the whole annotation process from subcorpus extraction to merging individual annotations, and allows for underspecified tags as well as tags beyond the sentence boundary and below the word boundary.

1 Introduction

We present the SALSA annotation tool, which offers a graphical environment for the manual annotation of frame-semantic roles (Baker et al., 1998) on top of syntactic trees in the TIGER format (Mengel and Lezius, 2000).

Recently there has been a growing interest in semantically annotated corpora, especially corpora annotated with semantic roles (Baker et al., 1998; Kingsbury et al., 2002; Hajičová, 1998). The hope is that statistical models that are trained on syntactically *and* semantically annotated corpora can improve upon existing systems on a broad range of tasks that can benefit from genuine semantic information, like question answering or semantic parsing.

However, good tools that support and speed up the annotation process are necessary for obtaining reasonably large annotated corpora. The SALSA annotation tool was designed for (and is used by) the SALSA project, the aim of which is to produce such a corpus for German (Erk et al., 2003).

2 Design

Input and output. The SALSA tool reads syntactically annotated corpora in the TIGER XML format (Mengel and Lezius, 2000). The list of frames that can be annotated is read directly from the FrameNet database of frames (see Section 3); new frames can also be added manually during annotation. Annotated corpora are stored in the SALSA/TIGER XML format, a modular extension of TIGER XML. This format keeps the semantic description layer separate from the syntactic layer, referring to nodes of the syntactic tree via unique node identifiers.

Annotation. Semantic information can be annotated in the form of flat trees on top of an existing syntactic graph. The main idea of the SALSA tool is to speed up the annotation of semantic roles by its point-and-click architecture and by letting annotators refer to the existing syntactic structure (see Sections 4 and 6). For the same purpose, we are planning to add support of an interactive semi-automatic annotation mode, similar to the *Annotate* tool (Plaehn and Brants, 2000).

Corpus management. In addition to the annotation process itself, the SALSA tool supports corpus management. Details are presented in Section 5.

Reusability. One design criterion for the tool was reusability for the annotation of other phenomena. This perspective is discussed in Section 7.

Implementation. The SALSA annotation tool was implemented in Java and uses the Swing library for its GUI. This operating system-independent design allows the application to be run on any platform with a recent Java runtime environment (<1.3). We have tested it successfully

under Windows, Linux, SunOS and Mac OS X.

3 FrameNet

The Berkeley FrameNet project (Baker et al., 1998) is based on Fillmore's Frame Semantics. A *frame* is a conceptual structure describing a situation. It is introduced by a *target* that can be a verb, noun or adjective. The roles, called *frame elements* (FEs), are local to frames and represent the agents and objects involved in that particular situation. For example, in "John bought a coat", "bought" introduces the frame COMMERCIAL_TRANSACTION with "John" in the role of the BUYER and "a coat" as the GOODS. The FrameNet project is constructing an XML database of frames along with a lexicon database that covers the core lexicon of English.

4 An example annotation

Figure 1 shows a screenshot of the SALSA tool, displaying a sentence annotated with two frames. Here, as in the rest of the paper, we draw all examples from the TIGER corpus (Brants et al., 2002) of German newspaper text.

Syntactic structure. The sentence in Figure 1, shown as terminals of the syntactic tree as well as in plain text format, is "Chaplins Sohn bewohnt noch das Haus der Familie und verkauft Videos in Vevey." (*Chaplin's son is still living in the family home and sells videos in Vevey.*) The syntactic structure of the sentence is shown as a tree with straight edges. The node labels (shown as dark circles) give the syntactic categories of constituents while the edge labels (shown as light squares) name the syntactic functions. For example the constituent "Chaplins Sohn" is tagged as an *NP* and as a subject (*SB*).

Frame representation. Frames are represented as trees of depth 1. The root node of a frame tree is labelled by the frame name, and the edges are labelled by the frame element names. The leaves of frame trees are nodes of the syntactic structure.

In the example sentence in Figure 1 two frames have been annotated. The verb "bewohnt" (*lives in*) evokes the frame INHABIT, with 2 frame elements: The subject-NP "Chaplins Sohn" (*Chaplin's son*) is the RESIDENT and the direct object is the LOCATION . The verb "verkauft" (*sells*) evokes the frame COMMERCIAL_TRANSACTION. Its frame elements SELLER and GOODS are assigned to "Chaplins Sohn" (*Chaplin's son*) and "Videos" (*videos*) respectively. Unassigned frame elements like CO_RESIDENT and MEANS just "hang off" the frame tree.

Frame annotation. The SALSA tool supports frame annotation in an intuitive and comfortable way. Frames are introduced by right-clicking on terminal words. A pop up menu lets the user choose a pre-selected frame or to create a new frame, as shown in Figure 1. Unassigned frame elements are assigned by simply dragging them to the appropriate node in the syntactic tree.

5 Corpus management

The SALSA tool supports an annotation process in which subcorpora are extracted from one large corpus, distributed to several annotators who tag them independently, and finally collected and merged into a single authoritative copy.

Creation of subcorpora. Subcorpora are created via an interface to TIGERSearch (Lezius, 2002), a search engine for treebanks. During subcorpus creation the user chooses frames from the FrameNet database that are appropriate for the subcorpus. These are the frames that are offered in the "Invoke frame" menu during annotation (cf. Figure 1). Additional frames can be added to the list during the annotation process.

Distribution of subcorpora. Newly created subcorpora are then distributed to annotators – each annotator gets an individual copy of the subcorpus file. After the annotation, the annotators move finished subcorpora to their "out" tray, where they can be collected by an administrator. After the subcorpus has been collected, it is no longer possible for the annotator to make modifications.

Merging of subcorpora. The SALSA tool provides a mode in which an administrator can create a single authoritative copy from two different annotated instances of the same subcorpus. In this mode, the two instances are merged automatically and differences are highlighted, such that the administrator can choose between the alternative annotations.

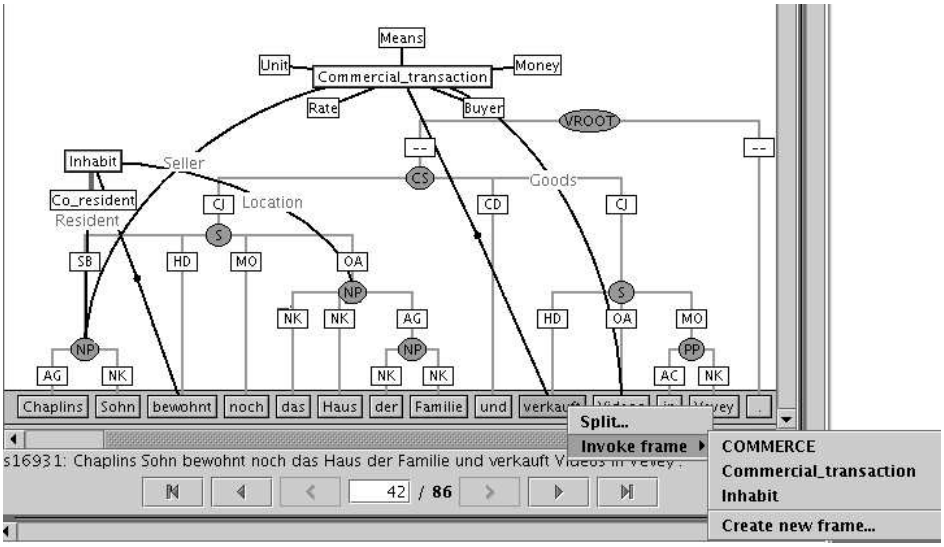

Figure 1: Frame Annotation.

6 Annotation features

The tool possesses a number of special features which are necessary for semantic role annotation.

Sentence context. Often the meaning of a sentence only becomes clear when the context is considered. In the SALSA tool, arbitrarily many sentences of context (preceding as well as succeeding) are shown on demand. Context can not only be viewed, it may also contain frame elements, which can be tagged with the tool. This happens e.g. in COMMUNICATION frames, where MESSAGE frame elements often span more than one sentence.

Discontinuous frame elements and targets. A target or a frame element may consist of more than one node of the syntactic structure. This happens for example with targets that have a separable verb prefix, as in "forderte . . . auf" (*demanded*). In such cases, annotators may split edges pointing to frame elements or targets.

Compound nouns. In a compound noun (written as a single word in German), different parts can realise different frame elements, or a frame element and a target. For example "Gagenforderung", (*demand for wages*) comprises both a REQUEST target and its MESSAGE. The SALSA tool lets annotators split a word into arbitrary parts and tag them separately.

Sentence tags. Sentence may be assigned tags that signal the need for later re-tagging (e.g. for coreference), mark a sentence as interesting, or indicate that a sentence does not belong in the current subcorpus.[1]

Underspecification. In word sense annotation it is not always possible to assign a single sense unequivocally. *Underspecified* assignment of a *set* of sense tags has been proposed as a solution (Kilgarriff and Rosenzweig, 2000; Fellbaum et al., 2001). In frame annotation, we find the same problem on *two* levels: for the choice of a frame for a target as well as for the assignment of frame elements to phrases. As an example of frame element underspecification, in "Machkämpfe" (*struggle for power*), "kämpfe" evokes HOSTILE ENCOUNTER, but it is not clear whether "Macht" should be the ISSUE fought over or the GOAL that the opponents hope to attain.

The SALSA tool allows underspecified assignment of two or more frames to the same target, two or more frame elements to the same constituent, underspecification with respect to a single frame element (indicating that it is not clear whether a certain constituent realises a frame element or not), and underspecification with regard to the extension of a frame element.

[1] This tag is necessary because of imperfect filters for subcorpus generation.

7 Reusability

Even though the tool was designed primarily for the task of semantic role annotation on the TIGER corpus, it can be used for other annotation tasks on other corpora, too.

Using other treebanks. The input format that is used by the SALSA tool is TIGER XML. It specifies trees with node and edge labels, allowing crossing edges, and secondary edges (used for conjunction and ellipsis in the TIGER corpus) that link arbitrary nodes of a tree. The terminals are labelled with lemma, part of speech and morphological information. This makes the TIGER XML format very powerful. Since it subsumes many other treebank formats, it should be possible to transform these corpora into the TIGER format.

Annotating other phenomena. The frame trees of the SALSA tool are trees of depth one with node and edge labels. They attach to arbitrary nodes of the syntactic tree, and they may span more than one sentence. This annotation scheme is general enough to accommodate completely different tasks, like annotating scope (with edge labels WIDE SCOPE and NARROW SCOPE) or coreference (with edge label ANTECEDENT).

8 Conclusion and outlook

We have presented the SALSA tool for the graphical annotation of (frame-)semantic roles on top of a syntactic structure given in TIGER format. It supports corpus management, provides underspecified tags, and offers role assignment beyond the sentence boundary and below the word boundary. Its open architecture makes it possible to use it for other annotation tasks.

The primary objective of the development of a graphical annotation tool was to speed up annotation compared to the former XML-based textual procedure. Since the tool has only been deployed recently, we do not have reliable figure yet, but we expect to gain a factor of 2 to 3. The next step is to extend the tool by an *interface to automatic role assignment systems*. Because training data is a major problem, as each frame has its own frame elements, we are currently using a rule-based learner that transfers annotations from a single corpus instance to other, similar instances. However, work on a more general statistical backend is under way. We expect that a transparent integration of automatic role assignment into the annotation process will yield another considerable acceleration.

Acknowledgements. The SALSA annotation tool was implemented by a team at CLT Sprachtechnologie GmbH under the direction of Daniel Bobbert, who we would also like to thank for valuable discussions.

References

C.F. Baker, C.J. Fillmore, and J.B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL*, Montreal, Canada.

S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

K. Erk, A. Kowalski, S. Padó, and M. Pinkal. 2003. Towards a resource for lexical semantics: A large German corpus with extensive semantic annotation. In *Proceedings of ACL-03*, Sapporo, Japan.

C. Fellbaum, M. Palmer, Hoa T. Dang, L. Delfs, and S. Wolf. 2001. Manual and automatic semantic annotation with WordNet. In *NAACL-2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh.

E. Hajičová. 1998. Prague Dependency Treebank: From Analytic to Tectogrammatical Annotation. In *Proceedings of TSD'98*, pages 45–50, Brno, Czech Republic.

A. Kilgarriff and J. Rosenzweig. 2000. Framework and results for English Senseval. *Computers and the Humanities*, 34(1-2).

P. Kingsbury, M. Palmer, and M. Marcus. 2002. Adding semantic annotation to the Penn TreeBank. In *Proceedings of HLT*.

W. Lezius. 2002. TIGERSearch - ein Suchwerkzeug für Baumbanken. In *Proceedings of Konvens 2002*, Saarbrücken, Germany.

A. Mengel and W. Lezius. 2000. An XML-based encoding format for syntactically annotated corpora. In *Proceedings of LREC-2000*, Athens, Greece.

O. Plaehn and T. Brants. 2000. Annotate - an efficient interactive annotation tool. In *Proceedings of ANLP-2000*, Seattle, WA.