# Frame assignment as word sense disambiguation

Katrin Erk
Department of Computational Linguistics
Saarland University
66123 Saarbrücken
erk@coli.uni-sb.de

December 1, 2004

**Introduction.** Recently there has been a growing interest in wide-coverage semantic analysis, especially with respect to semantic roles and word meaning, as a basis for inferences and for tasks like Question Answering. The Berkeley FrameNet database [1] is an interesting basis for such an analysis. It groups words and expressions (*lexical units*, *LUs* for short) into semantic classes (*frames*) and lists semantic roles (*frame elements*, *FEs* for short) for each class. A frame-semantic analysis of a text consists of two steps: the assignment of appropriate frames to LUs, and the assignment of (frame-specific) FEs. While there are many approaches to the automatic assignment of FEs, frame assignment has been less studied. Thompson et al [5] assign frames at the same time as FEs in an integrated approach. In this paper we treat frame assignment as a separate task, viewing it as word sense disambiguation. We train and test our system not on the FrameNet annotated examples, but on a frame-annotated corpus of German.

**The FrameNet corpus.** In the FrameNet database, frames are illustrated with example sentences from the British National Corpus. The most recent release of this corpus of example sentences comprises 68,118 annotated instances. While it has been used successfully as a training corpus for FE assignment, it is somewhat problematic as a training corpus for frame assignment. FrameNet is a growing resource, constructed frame by frame, which means that while each frame lists *all* LUs that introduce it, many LUs are still lacking some of their frames. For example, the verb *to have* has

| | FrameNet | SALSA |
|---|---|---|
| Percentage ambiguous LUs in the corpus | 10.7% | 43.4% |
| Average no. of frames per ambiguous LU | 2.27 | 4.13 |
| WSD baseline F-score: assign each LU its most frequent frame | 93% | 69.4% |

Table 1: Ambiguity of lexical units in the FrameNet corpus and the SALSA corpus snapshot

| | Precision | Recall | F-score |
|---|---|---|---|
| Without subcat features | 74.0 | 73.4 | 73.7 |
| With subcat features | 75.0 | 74.4 | 74.7 |

Table 2: Frame assignment as WSD: results (percent)

annotated examples only for the frame Birth ("to have a baby")[1]. The first column of Table 1 illustrates this sense coverage problem. This is a problem of the corpus, not of frame assignment as such, as our experiments with a frame-annotated corpus of German confirm.

**A snapshot of the SALSA corpus.** The SALSA project [3] is manually annotating the German TIGER corpus [2] with frames and FEs. Annotation proceeds lemma-wise rather than frame-wise, introducing additional lemma-specific proto-frames for instances not yet covered by FrameNet. For the WSD experiment we use a corpus snapshot containing 11,778 annotated instances, 11,220 verb instances and 558 noun instances. Of the two or more independent annotations of each lemma, one was chosen at random. The data may be expected to be rather noisy, as it consists of as yet unadjudicated annotations. The second column of Table 1 shows the degree of ambiguity in this corpus snapshot, which suggests that it does not suffer from the missing sense problem of the FrameNet corpus.

**Frame assignment as WSD.** We use a simple architecture for the system: a Naive Bayes classifier, which assigns each instance $x$ the sense $s$ that maximizes $P(s|x) = \frac{P(x|s)}{P(x)}P(s)$, with the simplifying (false) assumption that all features of the instance are independent: $P(x|s) = \prod_{f \text{ in } x} P(f|s)$. We use the rich feature set of Florian et al [4]: the bag-of-words context of the

[1] FrameNet also lists the frame Possession for *to have*, but there are no annotated example sentences for this sense of the verb.

target (words along with their lemmas and parts of speech), with a context window of only one sentence; local bigram and trigram collocations around the (main) target word; and head words and prepositions of complements and adjuncts of the target. In addition, we used features describing the subcat frame of the (main) target word. All features were weighted with a fixed weight specific to the feature type (determined empirically on a development set); context words were additionally weighted by their distance to the target. Then feature probabilities $P(f)$ and $P(f|s)$ were estimated using (smoothed) maximum likelihood estimation on the weighted features.

**Results on the SALSA corpus snapshot.** Table 2 shows the results for the system: Integration of subcategorization information improves the F-score by one point (which may be due to the fact that the targets were mainly verbs). One main difference between frame assignment as a single task and frame and FE assignment integrated is that potential FE head words are known in the latter approach. To explore the question of which approach will work better, we substituted frame element head words for the complement and adjunct head words. The result was almost unchanged, with an F-score of 74.4% (with subcat features).

## References


[1] C. Baker, C. Fillmore, and J. Lowe. The Berkeley FrameNet Project. In *Proceedings of ACL-98*, Montreal, Canada, 1998.

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

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

[4] R. Florian, S. Cucerzan, C. Schafer, and D. Yarowsky. Combining classifiers for word sense disambiguation. *Journal of Natural Language Engineering*, 8(4):327–431, 2002.

[5] C. Thompson, R. Levy, and C. Manning. A generative model for FrameNet semantic role labeling. In *Proceedings of the 14th ECML*, Croatia, 2003.