Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation

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

Abstract

In this paper, we present the results of an experiment in which we assess the usefulness of partial semi-automatic annotation for frame labeling. While we found no conclusive evidence that it can speed up human annotation, automatic pre-annotation does increase its overall quality.

1 Introduction

Linguistically annotated resources play a crucial role in natural language processing. Many recent advances in areas such as part-of-speech tagging, parsing, co-reference resolution, and semantic role labeling have only been possible because of the creation of manually annotated corpora, which then serve as training data for machine-learning based NLP tools. However, human annotation of linguistic categories is time-consuming and expensive. While this is already a problem for major languages like English, it is an even bigger problem for less-used languages.

This data acquisition bottleneck is a well-known problem and there have been numerous efforts to address it on the algorithmic side. Examples include the development of weakly supervised learning methods such as co-training and active learning. However, addressing only the algorithmic side is not always possible and not always desirable in all scenarios. First, some machine learning solutions are not as generally applicable or widely re-usable as one might think. It has been shown, for example, that co-training does not work well for problems which cannot easily be factorized into two independent views (Mueller et al., 2002; Ng and Cardie, 2003). Some active learning studies suggest both that the utility of the selected examples strongly depends on the model used for classification and that the example pool selected for one model can turn out to be sub-optimal when another model is trained on it at a later stage (Baldridge and Osborne, 2004). Furthermore, there are a number of scenarios for which there is simply no alternative to high-quality, manually annotated data; for example, if the annotated corpus is used for empirical research in linguistics (Meurers and Müller, 2007; Meurers, 2005).

In this paper, we look at this problem from the data creation side. Specifically we explore whether a semi-automatic annotation set-up in which a human expert corrects the output of an automatic system can help to speed up the annotation process without sacrificing annotation quality.

For our study, we explore the task of frame-semantic argument structure annotation (Baker et al., 1998). We chose this particular task because it is a rather complex – and therefore time-consuming – undertaking, and it involves making a number of different but interdependent annotation decisions for each instance to be labeled (e.g. frame assignment and labeling of frame elements, see Section 3.1). Semi-automatic support would thus be of real benefit.

More specifically, we explore the usefulness of automatic pre-annotation for the first step in the annotation process, namely frame assignment (word sense disambiguation). Since the available inventory of frame elements is dependent on the chosen frame, this step is crucial for the whole annotation process. Furthermore, semi-automatic annotation is more feasible for the frame labeling sub-task. Most automatic semantic role labeling systems (ASRL), including ours, tend to perform much better on frame assignment than on frame role labeling and correcting an erroneously chosen

Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009, pages 19–26,
Suntec, Singapore, 6-7 August 2009.

frame typically also requires fewer physical operations from the annotator than correcting a number of wrongly assigned frame elements.

We aim to answer three research questions in our study: First, we explore whether pre-annotation of frame labels can indeed speed up the annotation process. This question is important because frame assignment, in terms of physical operations of the annotator, is a relatively minor effort compared to frame role assignment and because checking a pre-annotated frame still involves all the usual mental operations that annotation from scratch does. Our second major question is whether annotation quality would remain acceptably high. Here the concern is that annotators might tend to simply go along with the pre-annotation, which would lead to an overall lower annotation quality than they could produce by annotating from scratch.[1] Depending on the purpose for which the annotations are to be used, trading off accuracy for speed may or may not be acceptable. Our third research question concerns the required quality of pre-annotation for it to have any positive effect. If the quality is too low, the annotation process might actually be slowed down because annotations by the automatic system would have to be deleted before the new correct one could be made. In fact, annotators might ignore the pre-annotations completely. To determine the effect of the pre-annotation quality, we not only compared a null condition of providing no prior annotation to one where we did, but we in fact compared the null condition to two different quality levels of pre-annotation, one that reflects the performance of a state-of-the-art ASRL system and an enhanced one that we artificially produced from the gold standard.

2 Related Work

While semi-automatic annotation is frequently employed to create labeled data more quickly (see, e.g., Brants and Plaehn (2000)), there are comparatively few studies which systematically look at the benefits or limitations of this approach. One of the earliest studies that investigated the advantages of manually correcting automatic annotations for linguistic data was carried out by Marcus et al. (1993) in the context of the construction of the Penn Treebank. Marcus et al. (1993) employed a post-correction set-up for both part-of-speech and syntactic structure annotation. For pos-tagging they compared the semi-automatic approach to a fully manual annotation. They found that the semi-automatic method resulted both in a significant reduction of annotation time, effectively doubling the word annotation rate, and in increased inter-annotator agreement and accuracy.

Chiou et al. (2001) explored the effect of automatic pre-annotation for treebank construction. For the automatic step, they experimented with two different parsers and found that both reduce overall annotation time significantly while preserving accuracy. Later experiments by Xue et al. (2002) confirmed these findings.

Ganchev et al. (2007) looked at semi-automatic gene identification in the biomedical domain. They, too, experimented with correcting the output of an automatic annotation system. However, rather than employing an off-the-shelf named entity tagger, they trained a tagger maximized for recall. The human annotators were then instructed to filter the annotation, rejecting falsely labeled expressions. Ganchev et al. (2007) report a noticeable increase in speed compared to a fully manual set-up.

The approach that is closest to ours is that of Chou et al. (2006) who investigate the effect of automatic pre-annotation for Propbank-style semantic argument structure labeling. However that study only looks into the properties of the semi-automatic set-up; the authors did not carry out a control study with a fully manual approach. Nevertheless Chou et al. (2006) provide an upper bound of the savings obtained by the semi-automatic process in terms of annotator operations. They report a reduction in annotation effort of up to 46%.

3 Experimental setup

3.1 Frame-Semantic Annotation

The annotation scheme we use is that of FrameNet (FN), a lexicographic project that produces a database of frame-semantic descriptions of English vocabulary. Frames are representations of prototypical events or states and their participants in the sense of Fillmore (1982). In the FN database, both frames and their participant roles are arranged in various hierarchical relations (most prominently, the is-a relation).

FrameNet links these descriptions of frames with the words and multi-words (lexical units, LUs) that evoke these conceptual structures. It also docu-

[1] This problem is also known in the context of resources that are collaboratively constructed via the web (Kruschwitz et al., 2009)

ments all the ways in which the semantic roles (frame elements, FEs) can be realized as syntactic arguments of each frame-evoking word by labeling corpus attestations. As a small example, consider the Collaboration frame, evoked in English by lexical units such as *collaborate*.v, *conspire*.v, *collaborator*.n and others. The core set of frame-specific roles that apply include Partner_1, Partner_2, Partners and Undertaking. A labeled example sentence is

(1) [The two researchers ^{Partners}] COLLABORATED [on many papers ^{Undertaking}].

FrameNet uses two modes of annotation: full-text, where the goal is to exhaustively annotate the running text of a document with all the different frames and roles that occur, and lexicographic, where only instances of particular target words used in particular frames are labeled.

3.2 Pilot Study

Prior to the present study we carried out a pilot experiment comparing manual and semi-automatic annotation of different segments of running text. In this experiment we saw no significant effect from pre-annotation. Instead we found that the annotation speed and accuracy depended largely on the order in which the texts were annotated and on the difficulty of the segments. The influence of order is due to the fact that FrameNet has more than 825 frames and each frame has around two to five core frame elements plus a number of non-core elements. Therefore even experienced annotators can benefit from the re-occuring of frames during the ongoing annotation process.

Drawing on our experiences with the first experiment, we chose a different experimental set-up for the present study. To reduce the training effect, we opted for annotation in lexicographic mode, restricting the number of lemmas (and thereby frames) to annotate, and we started the experiment with a training phase (see Section 3.5). Annotating in lexicographic mode also gave us better control over the difficulty of the different batches of data. Since these now consist of unrelated sentences, we can control the distribution of lemmas across the segments (see Section 3.4).

Furthermore, since the annotators in our pilot study had often ignored the error-prone pre-annotation, in particular for frame elements, we decided not to pre-annotate frame elements and to experiment with an enhanced level of pre-annotation to explore the effect of pre-annotation quality.

3.3 Annotation Set-Up

The annotators included the authors and three computational linguistics undergraduates who have been performing frame-semantic annotation for at least one year. While we use FrameNet data, our annotation set-up is different. The annotation consists of decorating automatically derived syntactic constituency trees with semantic role labels using the Salto tool (Burchardt et al., 2006) (see Figure 1). By contrast, in FrameNet annotation a chunk parser is used to provide phrase type and grammatical relations for the arguments of the target words. Further, FrameNet annotators need to correct mistakes of the automatic grammatical analysis, unlike in our experiment. The first annotation step, frame assignment, involves choosing the correct frame for the target lemma from a pull down menu; the second step, role assignment, requires the annotators to draw the available frame element links to the appropriate syntactic constituent(s).

The annotators performed their annotation on computers where access to the FrameNet website, where gold annotations could have been found, was blocked. They did, however, have access to local copies of the frame descriptions needed for the lexical units in our experiment. As the overall time needed for the annotation was too long to do in one sitting, the annotators did it over several days. They were instructed to record the time (in minutes) that they took for the annotation of each annotation session.

Our ASRL system for state-of-the-art pre-annotation was Shalmaneser (Erk and Pado, 2006). The enhanced pre-annotation was created by manually inserting errors into the gold standard.

3.4 Data

We annotated 360 sentences exemplifying all the senses that were defined for six different lemmas in FrameNet release 1.3. The lemmas were the verbs *rush, look, follow, throw, feel* and *scream*. These verbs were chosen for three reasons. First, they have enough annotated instances in the FN release that we could use some instances for testing and still be left with a set of instances sufficiently large to train our ASRL system. Second,we knew from prior work with our automatic role labeler that it had a reasonably good performance on these lemmas. Third, these LUs exhibit a range of difficulty in terms of the number of senses they have in FN (see Table 1) and the subtlety of the sense distinc-

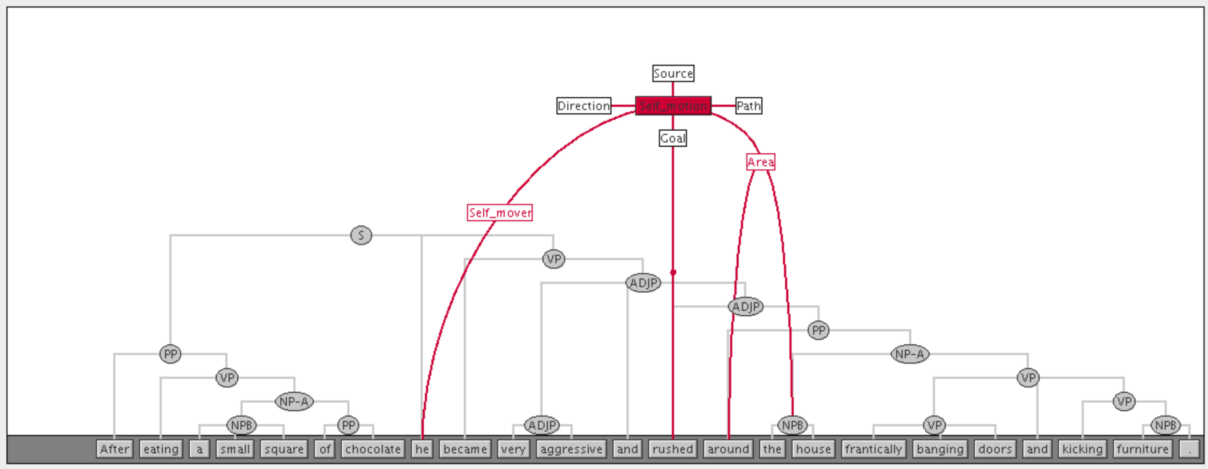

Figure 1: The Salto Annotation Tool

	Instances	Senses
feel	134	6
follow	113	3
look	185	4
rush	168	2
scream	148	2
throw	155	2

Table 1: Lemmas used

	1st	2nd	3rd	Annotators
Group I	E	S	N	5, 6
Group II	S	N	E	2, 4
Group III	N	E	S	1, 3

Table 2: Annotation condition by order and group

tions – e.g. the FrameNet senses of *look* are harder to distinguish than those of *rush*. We randomly grouped our sentences into three batches of equal size and for each batch we produced three versions corresponding to our three levels of annotation.

3.5 Study design

In line with the research questions that we want to address and the annotators that we have available, we choose an experimental design that is amenable to an analysis of variance. Specifically, we randomly assign our 6 annotators (1-6) to three groups of two (Groups I-III). Each annotator experiences all three annotation conditions, namely no pre-annotation (N), state-of-the-art pre-annotation (S), and enhanced pre-annotation (E). This is the within-subjects factor in our design, all other factors are between subjects. Namely, each group was randomly matched to one of three different orders in which the conditions can be experienced (see Table 2). The orderings are designed to control for the effects that increasing experience may have on speed and quality. While all annotators end up labeling all the same data, the groups also differ as to which batch of data is presented in which condition. This is intended as a check on any inherent differences in annotation difficulty that might exist between the data sets. Finally, to rule out difficulties with unfamiliar frames and frame elements needed for the lexical units used in this study, we provided some training to the annotators. In the week prior to the experiment, they were given 240 sentences exemplifying all 6 verbs in all their senses to annotate and then met to discuss any questions they might have about frame or FE distinctions etc. These 240 sentences were also used to train the ASRL system.

4 Results

In addition to time, we measured precision, recall and f-score for frame assignment and semantic role assignment for each annotator. We then performed an analysis of variance (ANOVA) on the outcomes of our experiment. Our basic results are presented in Table 3. As can be seen and as we expected, our annotators differed in their performance both with regard to annotation quality and speed. Below we discuss our results with respect to the research questions named above.

4.1 Can pre-annotation of frame assignment speed up the annotation process?

Not surprisingly, there are considerable differences in speed between the six annotators (Table 3),

Precision		Recall		F	t	p
Annotator 1						
94/103	91.3	94/109	86.2	88.68	75	N
99/107	92.5	99/112	88.4	90.40	61	E
105/111	94.6	105/109	96.3	95.44	65	S
Annotator 2						
93/105	88.6	93/112	83.0	85.71	135	S
86/98	87.8	86/112	76.8	81.93	103	N
98/106	92.5	98/113	86.7	89.51	69	E
Annotator 3						
95/107	88.8	95/112	84.8	86.75	168	N
103/110	93.6	103/112	92.0	92.79	94	E
99/113	87.6	99/113	87.6	87.60	117	S
Annotator 4						
106/111	95.5	106/112	94.6	95.05	80	S
99/108	91.7	99/113	87.6	89.60	59	N
105/112	93.8	105/113	92.9	93.35	52	E
Annotator 5						
104/110	94.5	(104/112)	92.9	93.69	170	E
91/103	88.3	(91/113)	80.5	84.22	105	S
96/100	96.0	(96/113)	85.0	90.17	105	N
Annotator 6						
102/106	96.2	102/112	91.1	93.58	124	E
94/105	89.5	94/112	83.9	86.61	125	S
93/100	93.0	93/113	82.3	87.32	135	N

Table 3: Results for frame assignment: precision, recall, f-score (F), time (t) (frame and role assignment), pre-annotation (p): **N**on, **E**nhanced, **S**halmaneser

which are statistically significant with $p \leq 0.05$. Focussing on the order in which the text segments were given to the annotators, we observe a significant difference ($p \leq 0.05$) in annotation time needed for each of the segments. With one exception, all annotators took the most time on the text segment given to them first, which hints at an ongoing training effect.

The different conditions of pre-annotation (none, state-of-the-art, enhanced) did not have a significant effect on annotation time. However, all annotators except one were in fact faster under the enhanced condition than under the unannotated condition. The one annotator who was not faster annotated the segment with the enhanced pre-annotation before the other two segments; hence there might have been an interaction between time savings from pre-annotation and time savings due to a training effect. This interaction between training effect and degree of pre-annotation might be one reason why we do not find a significant effect between annotation time and pre-annotation condition. Another reason might be that the pre-annotation only reduces the physical effort needed to *annotate* the correct frame which is relatively minor compared to the cognitive effort of *determining* (or verifying) the right frame, which is required for all degrees of pre-annotation.

4.2 Is annotation quality influenced by automatic pre-annotation?

To answer the second question, we looked at the relation between pre-annotation condition and f-score. Even though the results in f-score for the different annotators vary in extent (Table 4), there is no significant difference between annotation quality for the six annotators.

Anot1	Anot2	Anot3	Anot4	Anot5	Anot6
91.5	85.7	89.0	92.7	89.4	89.2

Table 4: Average f-score for the 6 annotators

Next we performed a two-way ANOVA (Within-Subjects design), and crossed the dependent variable (f-score) with the two independent variables (order of text segments, condition of pre-annotation). Here we found a significant effect ($p \leq 0.05$) for the impact of pre-annotation on annotation quality. All annotators achieved higher f-scores for frame assignment on the enhanced pre-annotated text segments than on the ones with no pre-annotation. With one exception, all annotators also improved on the already high baseline for the enhanced pre-annotation (Table 5).

Seg.	Precision		Recall	f-score
Shalmaneser				
A	(70/112)	62.5	(70/96)	72.9	67.30
B	(75/113)	66.4	(75/101)	74.3	70.13
C	(66/113)	58.4	(66/98)	67.3	62.53
Enhanced Pre-Annotation					
A	(104/112)	92.9	(104/111)	93.7	93.30
B	(103/112)	92.0	(103/112)	92.0	92.00
C	(99/113)	87.6	(99/113)	87.6	87.60

Table 5: Baselines for automatic pre-annotation (Shalmaneser) and enhanced pre-annotation

The next issue concerns the question of whether annotators make different types of errors when provided with the different styles of pre-annotation. We would like to know if erroneous frame assignment, as done by a state-of-the-art ASRL will tempt annotators to accept errors they would not make in the first place. To investigate this issue, we compared f-scores for each of the frames for all three pre-annotation conditions with f-scores for frame assignment achieved by Shalmaneser. The boxplot in Figure 2 shows the distribution of f-scores for each frame for the different pre-annotation styles and for Shalmaneser. We can see that the same

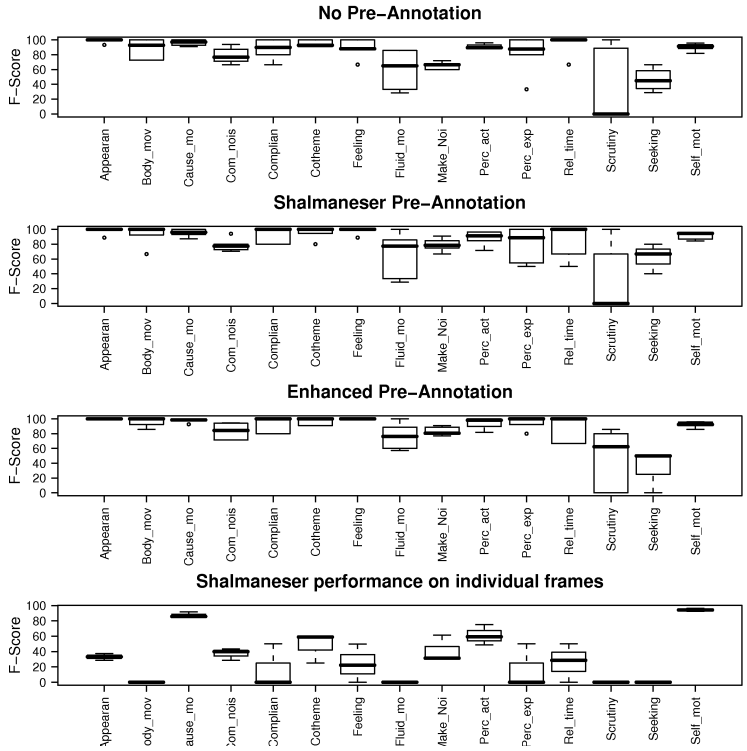

Figure 2: F-Scores per frame for human annotators on different levels of pre-annotation and for Shalmaneser

error types are made by human annotators throughout all three annotation trials, and that these errors are different from the ones made by the ASRL.

Indicated by f-score, the most difficult frames in our data set are Scrutiny, Fluidic_motion, Seeking, Make_noise and Communication_noise. This shows that automatic pre-annotation, even if noisy and of low quality, does not corrupt human annotators on a grand scale. Furthermore, if the pre-annotation is good it can even improve the overall annotation quality. This is in line with previous studies for other annotation tasks (Marcus et al., 1993).

4.3 How good does pre-annotation need to be to have a positive effect?

Comparing annotation quality on the automatically pre-annotated texts using Shalmaneser, four out of six annotators achieved a higher f-score than on the non-annotated sentences. The effect, however, is not statistically significant. This means that pre-annotation produced by a state-of-the-art ASRL system is not yet good enough a) to significantly speed up the annotation process, and b) to improve the quality of the annotation itself. On the positive side, we also found no evidence that the error-prone pre-annotation decreases annotation quality.

Most interestingly, the two annotators who showed a decrease in f-score on the text segments pre-annotated by Shalmaneser (compared to the text segments with no pre-annotation provided) had been assigned to the same group (Group I). Both had first annotated the enhanced, high-quality pre-annotation, in the second trial the sentences pre-annotated by Shalmaneser, and finally the texts with no pre-annotation. It might be possible that they benefitted from the ongoing training, resulting in a higher f-score for the third text segment (no pre-annotation). For this reason, we excluded their annotation results from the data set and performed another ANOVA, considering the remaining four annotators only.

Figure 3 illustrates a noticeable trend for the interaction between pre-annotation and annotation quality: all four annotators show a decrease in annotation quality on the text segments without pre-annotation, while both types of pre-annotation (Shalmaneser, Enhanced) increase f-scores for human annotation. There are, however, differences between the impact of the two pre-annotation types on human annotation quality: two annotators show better results on the enhanced, high-quality pre-

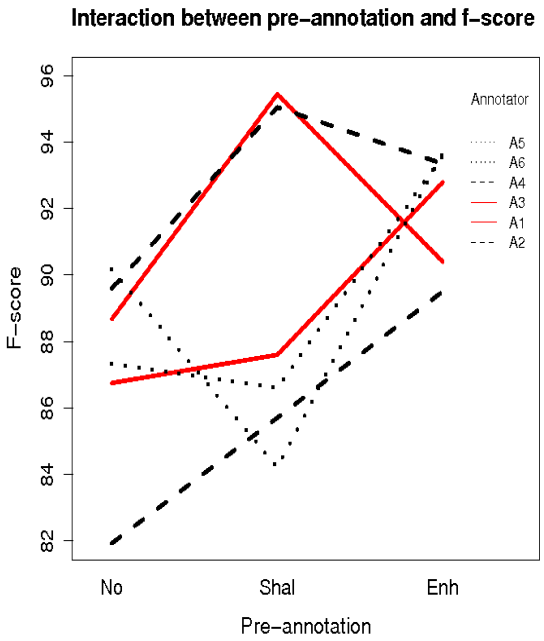

Figure 3: Interaction between pre-annotation and f-score

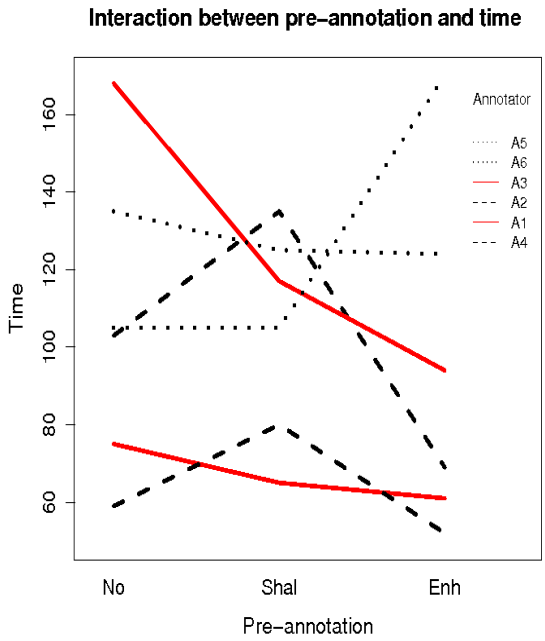

Figure 4: Interaction between pre-annotation and time

annotation, the other two perform better on the texts pre-annotated by the state-of-the-art ASRL. The interaction between pre-annotation and f-score computed for the four annotators is weakly significant with $p \leq 0.1$.

Next we investigated the influence of pre-annotation style on annotation time for the four annotators. Again we can see an interesting pattern: The two annotators (A1, A3) who annotated in the order N-E-S, both take most time for the texts without pre-annotation, getting faster on the text pre-processed by Shalmaneser, while the least amount of time was needed for the enhanced pre-annotated texts (Figure 4). The two annotators (A2, A4) who processed the texts in the order S-N-E, showed a continuous reduction in annotation time, probably caused by the interaction of training and data quality. These observations, however, should be taken with a grain of salt, as they outline trends, but due to the low number of annotators, could not be substantiated by statistical tests.

4.4 Semantic Role Assignment

As described in Section 3.5, we provided pre-annotation for frame assignment only, therefore we did not expect any significant effects of the different conditions of pre-annotation on the task of semantic role labeling. To allow for a meaningful comparison, the evaluation of semantic role assignment was done on the subset of frames annotated correctly by all annotators.

As with frame assignment, there are considerable differences in annotation quality between the annotators. In contrast to frame assignment, here the differences are statistically significant ($p \leq 0.05$). Table 6 shows the average f-score for each annotator on the semantic role assignment task.

Anot1	Anot2	Anot3	Anot4	Anot5	Anot6
85.2	80.1	87.7	89.2	82.5	84.3

Table 6: Average f-scores for the 6 annotators

As expected, neither the condition of pre-annotation nor the order of text segments had any significant effect on the quality of semantic role assignment.[2]

5 Conclusion and future work

In the paper we presented experiments to assess the benefits of partial automatic pre-annotation on a frame assignment (word sense disambiguation) task. We compared the impact of a) pre-annotations

[2]The annotation of frame and role assignment was done as a combined task, therefore we do not report separate results for annotation time for semantic role assignment.

provided by a state-of-the-art ASRL, and b) enhanced, high-quality pre-annotation on the annotation process. We showed that pre-annotation has a positive effect on the quality of human annotation: the enhanced pre-annotation clearly increased f-scores for all annotators, and even the noisy, error-prone pre-annotations provided by the ASRL system did not lower the quality of human annotation.

We suspect that there is a strong interaction between the order in which the text segments are given to the annotators and the three annotation conditions, resulting in lower f-scores for the group of annotators who processed the ASRL pre-annotations in the first trial, where they could not yet profit from the same amount of training as the other two groups.

The same problem occurs with annotation time. We have not been able to show that automatic pre-annotation speeds up the annotation process. However, we suspect that here, too, the interaction between training effect and annotation condition made it difficult to reach a significant improvement. One way to avoid the problem would be a further split of the test data, so that the different types of pre-annotation could be presented to the annotators at different stages of the annotation process. This would allow us to control for the strong bias through incremental training, which we cannot avoid if one group of annotators is assigned data of a given pre-annotation type in the first trial, while another group encounters the same type of data in the last trial. Due to the limited number of annotators we had at our disposal as well as the amount of time needed for the experiments we could not sort out the interaction between order and annotation conditions. We will take this issue up in future work, which also needs to address the question of how good the automatic pre-annotation should be to support human annotation. F-scores for the enhanced pre-annotation provided in our experiments were quite high, but it is possible that a similar effect could be reached with automatic pre-annotations of somewhat lower quality.

The outcome of our experiments provides strong motivation to improve ASRL systems, as automatic pre-annotation of acceptable quality does increase the quality of human annotation.

References

C. F. Baker, C. J. Fillmore, J. B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics*, 86–90, Morristown, NJ, USA. Association for Computational Linguistics.

J. Baldridge, M. Osborne. 2004. Active learning and the total cost of annotation. In *Proceedings of EMNLP*.

T. Brants, O. Plaehn. 2000. Interactive corpus annotation. In *Proceedings of LREC-2000*.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padó. 2006. SALTO – a versatile multi-level annotation tool. In *Proceedings of LREC*.

F.-D. Chiou, D. Chiang, M. Palmer. 2001. Facilitating treebank annotation using a statistical parser. In *Proceedings of HLT-2001*.

W.-C. Chou, R. T.-H. Tsai, Y.-S. Su, W. Ku, T.-Y. Sung, W.-L. Hsu. 2006. A semi-automatic method for annotation a biomedical proposition bank. In *Proceedings of FLAC-2006*.

K. Erk, S. Pado. 2006. Shalmaneser - a flexible toolbox for semantic role assignment. In *Proceedings of LREC*, Genoa, Italy.

C. J. F. C. J. Fillmore, 1982. *Linguistics in the Morning Calm*, chapter Frame Semantics, 111–137. Hanshin Publishing, Seoul, 1982.

K. Ganchev, F. Pereira, M. Mandel, S. Carroll, P. White. 2007. Semi-automated named entity annotation. In *Proceedings of the Linguistic Annotation Workshop*, 53–56, Prague, Czech Republic. Association for Computational Linguistics.

U. Kruschwitz, J. Chamberlain, M. Poesio. 2009. (linguistic) science through web collaboration in the ANAWIKI project. In *Proceedings of WebSci'09*.

M. P. Marcus, B. Santorini, M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

W. D. Meurers, S. Müller. 2007. Corpora and syntax (article 44). In A. Lüdeling, M. Kytö, eds., *Corpus linguistics*. Mouton de Gruyter, Berlin.

W. D. Meurers. 2005. On the use of electronic corpora for theoretical linguistics. case studies from the syntax of german. *Lingua*, 115(11):1619–1639. `http://purl.org/net/dm/papers/meurers-03.html`.

C. Mueller, S. Rapp, M. Strube. 2002. Applying co-training to reference resolution. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, 352–359, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

V. Ng, C. Cardie. 2003. Bootstrapping coreference classifiers with multiple machine learning algorithms. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*.

N. Xue, F.-D. Chiou, M. Palmer. 2002. Building a large-scale annotated chinese corpus. In *Proceedings of Coling-2002*.