Bringing Active Learning to Life

Ines Rehbein
Computational Linguistics
Saarland University

Josef Ruppenhofer
Computational Linguistics
Saarland University

Alexis Palmer
Computational Linguistics
Saarland University

{rehbein|josefr|apalmer}@coli.uni-sb.de

Abstract

Active learning has been applied to different NLP tasks, with the aim of limiting the amount of time and cost for human annotation. Most studies on active learning have only simulated the annotation scenario, using prelabelled gold standard data. We present the first active learning experiment for Word Sense Disambiguation with human annotators in a realistic environment, using fine-grained sense distinctions, and investigate whether AL can reduce annotation cost and boost classifier performance when applied to a real-world task.

1 Introduction

Active learning has recently attracted attention as having the potential to overcome the knowledge acquisition bottleneck by limiting the amount of human annotation needed to create training data for statistical classifiers. Active learning has been shown, for a number of different NLP tasks, to reduce the number of manually annotated instances needed for obtaining a consistent classifier performance (Hwa, 2004; Chen et al., 2006; Tomanek et al., 2007; Reichart et al., 2008).

The majority of such results have been achieved by simulating the annotation scenario using prelabelled gold standard annotations as a stand-in for real-time human annotation. Simulating annotation allows one to test different parameter settings without incurring the cost of human annotation. There is, however, a major drawback: we do not know whether the results of experiments performed using hand-corrected data carry over to real-world scenarios in which individual human annotators produce noisy annotations. In addition, we do not know to what extent error-prone annotations mislead the learning process. A systematic study of the impact of erroneous annotation on classifier performance in an active learning (AL) setting is overdue. We need to know a) whether the AL approach can really improve classifier performance and save annotation time when applied in a real-world scenario with noisy data, and b) whether AL works for classification tasks with fine-grained or complex annotation schemes and a low inter-annotator agreement.

In this paper we bring active learning to life in the context of frame semantic annotation of German texts within the SALSA project (Burchardt et al., 2006). Specifically, we apply AL methods for learning to assign semantic frames to predicates, following Erk (2005) in treating frame assignment as a Word Sense Disambiguation task. Under our fine-grained annotation scheme, annotators have to deal with a high level of ambiguity, resulting in low inter-annotator agreement for some word senses. This fact, along with the potential for wrong annotation decisions or possible biases from individual annotators, results in an annotation environment in which we get noisy data which might mislead the classifier. A second characteristic of our scenario is that there is no gold standard for the newly annotated data, which means that evaluation is not straightforward. Finally, we have multiple annotators whose deci-

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 949–957,
Beijing, August 2010

sions on particular instances may diverge, raising the question of which annotations should be used to guide the AL process. This paper thus investigates whether active learning can be successfully applied in a real-world scenario with the particular challenges described above.

Section 2 of the paper gives a short overview of the AL paradigm and some related work, and Section 3 discusses the multi-annotator scenario. In Section 4 we present our experimental design and describe the data we use. Section 5 presents results, and Section 6 concludes.

2 Active Learning

The active learning approach aims to reduce the amount of manual annotation needed to create training data sufficient for developing a classifier with a given performance. At each iteration of the AL cycle, the actual knowledge state of the learner guides the learning process by determining which instances are chosen next for annotation. The main goal is to advance the learning process by selecting instances which provide important information for the machine learner.

In a typical active learning scenario, a small set of manually labelled seed data serves as the initial training set for the classifier (learner). Based on the predictions of the classifier, a large pool of unannotated instances is queried for the next instance (or batch of instances) to be presented to the human annotator (sometimes called the oracle). The underlying active learning algorithm controlling the learning process tries to select the most informative instances in order to get a strong boost in classifier performance. Different methods can be used for determining informativity of instances. We use uncertainty sampling (Cohn et al., 1995) in which “most informative” instances are those for which the classifier has the lowest confidence in its label predictions. The rough intuition behind this selection method is that it identifies instance types which have yet to be encountered by the classifier. The learning process proceeds by presenting the selected instances to the human annotator, who assigns the correct label. The newly-annotated instances are added to the seed data and the classifier is re-trained on the new data set. The newly trained classifier now picks the next instances, based on its updated knowledge, and the process repeats. If the learning process can provide precisely that information which the classifier still needs to learn, a smaller number of instances should suffice to achieve the same accuracy as on a larger training set of randomly selected training examples.

Active learning has been applied to a number of natural language processing tasks like POS tagging (Ringger et al., 2007), NER (Laws and Schütze, 2008; Tomanek and Hahn, 2009), syntactic parsing (Osborne and Baldridge, 2004; Hwa, 2004), Word Sense Disambiguation (Chen et al., 2006; Chan and Ng, 2007; Zhu and Hovy, 2007; Zhu et al., 2008) and morpheme glossing for language documentation (Baldridge and Palmer, 2009). While most of these studies successfully show that the same classification accuracy can be achieved with a substantially smaller data set, these findings are mostly based on simulations using gold standard data.

For our task of Word Sense Disambiguation (WSD), mixed results have been achieved. AL seems to improve results in a WSD task with coarse-grained sense distinctions (Chan and Ng, 2007), but the results of (Dang, 2004) raise doubts as to whether AL can successfully be applied to a fine-grained annotation scheme, where Inter-Annotator Agreement (IAA) is low and thus the consistency of the human annotations decreases. In general, AL has been shown to reduce the cost of annotation when applied to classification tasks where a single human annotator predicts labels for new data points with a reasonable consistency and accuracy. It is not clear whether the same settings can be applied to a multi-annotator environment where IAA is low.

3 Active Learning in a realistic task including multiple annotators

Another possible difference between active learning simulations and real-world scenarios is the multi-annotator environment. In such a setting, two or more annotators assign labels to the same instances, which are then merged to check for conflicting decisions from different annotators. This is standard practise in many annotation projects doing fine-grained semantic annotation with a high level of ambiguity, and it necessitates that all annotators work on the same data set.

Replicating an active learning simulation on hand-corrected data, starting with a fixed set of seed data and fixed parameter settings, using the same algorithm, will always result in the same training set selected from the pool. Human annotators, however, will assign different labels to the same instances, thus influencing the selection of the next instance from the pool. This means that individual annotators might end up with very different sets of annotated data, depending on factors like their interpretation of the annotation guidelines, an implicit bias towards a particular label, or simply errors made during annotation.

There is not much work addressing this problem. (Donmez and Carbonell, 2008) consider modifications of active learning to accommodate variability of annotators. (Baldridge and Palmer, 2009) present a real-world study with human annotators in the context of language documentation. The task consists of producing interlinear glossed text, including morphological and grammatical analysis, and can be described as a sequence labelling task. Annotation cost is measured as the actual time needed for annotation. Among other settings, the authors compare the performance of two annotators with different grades of expertise. The classifier trained on the data set created by the expert annotator in an active learning setting does obtain a higher accuracy on the gold standard. For the non-expert annotator, however, the active learning setting resulted in a lower accuracy than for a classifier trained on a randomly selected data set. This finding suggests that the quality of annotation needs to be high enough for active learning to actually work, and that annotation noise is a problem for AL.

There are two problems arising from this:

1. It is not clear whether active learning will work when applied to noisy data
2. It is not straightforward to apply active learning to a real-world scenario, where low IAA asks for multiple annotators

In our experiment we address these questions by systematically investigating the impact of annotation noise on classifier performance and on the composition of the training set. The next section presents the experimental design and the data used in our experiment.

4 Experimental Design

In the experiment we annotated 8 German causation nouns, namely *Ausgang, Anlass, Ergebnis, Resultat, Grund, Konsequenz, Motiv, Quelle* (outcome, occasion, effect, result, reason, consequence, motive, source of experience). These nouns were chosen because they exhibit a range of difficulty in terms of the number of senses they have in our annotation scheme. They all encode subtle distinctions between different word senses, but some of them are clearly easier to disambiguate than others. For instance, although *Ausgang* has 9 senses, they are easier to distinguish for humans than the 4 senses of *Konsequenz*.

Six annotators participated in the experiment. While all annotators were trained, having at least one year experience in frame-semantic annotation, one of the annotators is an expert with several years of training and working experience in the Berkeley FrameNet Project. This annotator also defined the frames (word senses) used in our experiment.

Prior to the experiment, all annotators were given 100 randomly chosen sentences. After annotating the training data, problematic cases were discussed to make sure that the annotators were familiar with the fine-grained distinctions between word senses in the annotation scheme. The data sets used for training were adjudicated by two of the annotators (one of them being the expert) and then used as a gold standard to test classifier performance in the active learning process.

4.1 Data and Setup

For each lemma we extracted sentences from the Wahrig corpus[1] containing this particular lemma. The annotators had to assign word senses to 300 instances for each target word, split into 6 packages of 50 sentences each. This resulted in 2,400 annotated instances per annotator (14,400 annotated instances in total). The annotation was done

[1] The Wahrig corpus includes more than 113 mio. sentences from German newspapers and magazines covering topics such as politics, science, fashion, and others.

Anlass	Motiv	Konsequenz	Quelle	Ergebnis / Resultat	Ausgang	Grund
Occasion (37)	Motif (47)	Causation (32)	Relational_nat_feat.(3)	Causation (4/10)	Outcome (67)	Causation (24)
Reason (63)	Reason(53)	Level_of_det.(6)	Source_of_getting (14)	Competitive_score(12/36)	Have_leave (4)	Reason (58)
		Response (61)	Source_of_exp. (14)	Decision (11/6)	Portal (21)	Death (1)
		MWE1 (1)	Source_of_info. (56)	Efficacy (2/3)	Outgoing_goods (4)	Part_orientation. (0)
			Well (6)	Finding_out (24/23)	Ostomy (0)	Locale_by_owner(3)
			Emissions_source (7)	Mathematics (1/0)	Origin (5)	Surface_earth (0)
				Operating_result (36/5)	Tech_output (7)	Bottom_layer (0)
				Outcome (10/17)	Process_end (2)	Soil (1)
					Departing (1)	CXN1 (0)
						CXN2 (0)
						MWE1 (0)
						MWE2 (10)
						MWE3 (0)
						MWE4 (3)
						MWE5 (0)
						MWE6 (0)
Fleiss' kappa for the 6 annotators for the 150 instances annotated in the random setting						
0.67	0.79	0.55	0.77	0.63 / 0.59	0.82	0.43

Table 1: 8 causation nouns and their word senses (numbers in brackets give the distribution of word senses in the gold standard (100 sentences); CXN: constructions, MWE: multi-word expressions; note that Ergebnis and Resultat are synonyms and therefore share the same set of frames.)

using a Graphical User Interface where the sentence was presented to the annotator, who could choose between all possible word senses listed in the GUI. The annotators could either select the frame by mouse click or use keyboard shortcuts. For each instance we recorded the time it took the annotator to assign an appropriate label. To ease the reading process the target word was highlighted.

As we want to compare time requirements needed for annotating random samples and sentences selected by active learning, we had to control for training effects which might speed up the annotation. Therefore we changed the annotation setting after each package, meaning that the first annotator started with 50 sentences randomly selected from the pool, then annotated 50 sentences selected by AL, followed by another 50 randomly chosen sentences, and so on. We divided the annotators into two groups of three annotators each. The first group started annotating in the random setting, the second group in the AL setting. The composition of the groups was changed for each lemma, so that each annotator experienced all different settings during the annotation process. The annotators were not aware of which setting they were in.

Pool data For the random setting we randomly selected three sets of sentences from the Wahrig corpus which were presented for annotation to all six annotators. This allows us to compare annotation time and inter-annotator agreement between the annotators. For the active learning setting we randomly selected three sets of 2000 sentences each, from which the classifier could pick new instances during the annotation process. This means that for each trial the algorithm could select 50 instances out of a pool of 2000 sentences. On any given AL trial each annotator uses the same pool as all the other annotators. In an AL simulation with fixed settings and gold standard labels this would result in the same subset of sentences selected by the classifier. For our human annotators, however, due to different annotation decisions the resulting set of sentences is expected to differ.

Sampling method Uncertainty sampling is a standard sampling method for AL where new instances are selected based on the confidence of the classifier for predicting the appropriate label. During early stages of the learning process when the classifier is trained on a very small seed data set, it is not beneficial to add the instances with the lowest classifier confidence. Instead, we use a dynamic version of uncertainty sampling (Rehbein and Ruppenhofer, 2010), based on the confidence of a maximum entropy classifier[2], taking into account how much the classifier has learned so far. In each iteration one new instance is selected from the pool and presented to the oracle. After annotation the classifier is retrained on the new data set. The modified uncertainty sampling results in a more robust classifier performance during early stages of the learning process.

[2] http://maxent.sourceforge.net

	Anlass		Motiv		Konsequenz		Quelle		Ergebnis		Resultat		Ausgang		Grund	
	R	U	R	U	R	U	R	U	R	U	R	U	R	U	R	U
A_1	8.6	9.6	5.9	6.6	10.7	10.5	6.0	4.8	10.5	7.4	10.1	9.6	6.4	10.0	10.2	11.1
A_2	4.4	5.7	4.8	5.9	8.2	9.2	4.9	4.9	6.4	4.4	11.7	8.5	5.1	7.7	9.0	9.3
A_3	9.9	9.2	6.8	6.7	6.8	8.3	7.4	6.1	9.4	7.6	9.0	12.3	7.5	8.5	11.7	10.2
A_4	5.8	4.9	3.6	3.6	9.9	11.3	4.8	3.5	7.9	7.1	9.7	11.1	3.6	4.1	9.9	9.4
A_5	3.0	3.5	3.0	2.6	4.8	4.9	3.8	3.0	6.8	4.8	6.7	6.1	3.1	3.5	6.3	6.0
A_6	5.4	6.3	5.3	4.7	6.7	8.6	5.4	4.6	7.8	6.1	8.7	9.0	6.9	6.6	9.3	8.5
ø	**6.2**	**6.5**	**4.9**	**5.0**	**7.8**	**8.8**	**5.4**	**4.5**	**8.1**	**6.2**	**9.3**	**9.4**	**5.4**	**6.7**	**9.4**	**9.1**
sl	25.8	27.8	27.8	26.0	24.2	25.8	24.9	26.5	25.7	25.2	29.0	35.9	25.5	27.9	26.8	29.7

Table 2: Annotation time (sec/instance) per target/annotator/setting and average sentence length (sl)

5 Results

The basic idea behind active learning is to select the most informative instances for annotation. The intuition behind "more informative" is that these instances support the learning process, so we might need fewer annotated instances to achieve a comparable classifier performance, which could decrease the cost of annotation. On the other hand, "more informative" also means that these instances might be more difficult to annotate, so it is only fair to assume that they might need more time for annotation, which increases annotation cost. To answer the question of whether AL reduces annotation cost or not we have to check a) how long it took the annotators to assign labels to the AL samples compared to the randomly selected instances, and b) how many instances we need to achieve the best (or a sufficient) performance in each setting. Furthermore, we want to investigate the impact of active learning on the distribution of the resulting training sets and study the correlation between the performance of the classifier trained on the annotated data and these factors: the difficulty of the annotation task (assessed by IAA), expertise and individual properties of the annotators.

5.1 Does AL speed up the annotation process when working with noisy data?

Table 2 reports annotation times for each annotator and target for random sampling (R) and uncertainty sampling (U). For 5 out of 8 targets the time needed for annotating in the AL setting (averaged over all annotators) was higher than for annotating the random samples. To investigate whether this might be due to the length of the sentences in the samples, Table 2 shows the average sentence length for random samples and AL samples for each target lemma. Overall, the sentences selected by the classifier during AL are longer (26.2 vs. 28.1 token per sentence), and thus may take the annotators more time to read.[3] However, we could not find a significant correlation (Spearman rank correlation test) between sentence length and annotation time, nor between sentence length and classifier confidence.

The three target lemmas which took longer to annotate in the random setting are *Ergebnis* (result), *Grund* (reason) and *Quelle* (source of experience). This observation cannot be explained by sentence length. While sentence length for *Ergebnis* is nearly the same in both settings, for *Grund* and *Quelle* the sentences picked by the classifier in the AL setting are significantly longer and therefore should have taken more time to annotate. To understand the underlying reason for this we have to take a closer look at the distribution of word senses in the data.

5.2 Distribution of word senses in the data

In the literature it has been stated that AL implicitly alleviates the class imbalance problem by extracting more balanced data sets, while random sampling tends to preserve the sense distribution present in the data (Ertekin et al., 2007). We could not replicate this finding when using noisy data to guide the learning process. Table 3 shows the distribution of word senses for the target lemma *Ergebnis* a) in the gold standard, b) in the random samples, and c) in the AL samples.

The variance in the distribution of word senses in the random samples and the gold standard can

[3] The correlation between sentence length and annotation time is not obvious, as the annotators only have to label one target in each sentence. For ambiguous sentences, however, reading time may be longer, while for the clear cases we do not expect a strong effect.

Ergebnis			
Frame	**gold (%)**	**R (%)**	**U (%)**
Causation	4.0	4.8	3.7
Outcome	10.0	17.8	10.5
Finding_out	24.0	26.2	8.2
Efficacy	2.0	0.8	0.1
Decision	11.0	5.1	3.2
Mathematics	1.0	1.6	0.4
Operating_result	36.0	24.5	66.7
Competitive_score	12.0	19.2	7.2

Table 3: Distribution of frames (word senses) for the lemma *Ergebnis* in the gold standard (100 sentences), in the random samples (R) and AL samples (U) (150 sentences each)

	Random			**Uncertainty**		
	50	100	150	50	100	150
Anlass	0.85	**0.86**	0.85	0.84	0.85	0.84
Motiv	0.57	0.62	0.63	0.64	0.67	**0.70**
Konseq.	0.55	0.59	0.60	0.61	**0.62**	**0.62**
Quelle	0.56	0.53	0.54	0.52	0.52	**0.57**
Ergebnis	0.39	**0.42**	0.41	0.39	0.37	0.38
Resultat	0.31	0.35	**0.37**	0.32	0.34	0.34
Ausgang	0.67	0.69	0.69	0.68	0.72	**0.74**
Grund	**0.48**	0.47	0.47	0.47	0.44	**0.48**

Table 4: Avg. classifier performance (acc.) over all annotators for the 8 target lemmas when training on 50, 100 and 150 annotated instances for random samples and uncertainty samples

be explained by low inter-annotator agreement caused by the high level of ambiguity for the target lemmas. The frame distribution in the data selected by uncertainty sampling, however, crucially deviates from those of the gold standard and the random samples. A disproportionately high 66% of the instances selected by the classifier have been assigned the label Operating_result by the human annotators. This is the more surprising as this frame is fairly easy for humans to distinguish.

The classifier, however, proved to have serious problems learning this particular word sense and thus repeatedly selected more instances of this frame for annotation. As a result, the distribution of word senses in the training set for the uncertainty samples is highly skewed, having a negative effect on the overall classifier performance. The high percentage of instances of the "easy-to-decide" frame Operating_result explains why the instances for *Ergebnis* took less time to annotate in the AL setting. Thus we can conclude that annotating the same number of instances on average takes more time in the AL setting, and that this effect is not due to sentence length.

5.3 What works, what doesn't, and why

For half of the target lemmas *(Motiv, Konsequenz, Quelle, Ausgang),* we did obtain best results in the AL setting (Table 4). For *Ausgang* and *Motiv* AL gives a substantial boost in classifier performance of 5% and 7% accuracy, while the gains for *Konsequenz* and *Quelle* are somewhat smaller with 2% and 1%, and for *Grund* the highest accuracy was reached on both the AL and the random sample.

Figure 1 (top row) shows the learning curves for *Resultat*, our worst-performing lemma, for the classifier trained on the manually annotated samples for each individual annotator. The solid black line represents the majority baseline, obtained by assigning the most frequent word sense in the gold standard to all instances. For both random and AL settings, results are only slightly above the baseline. The curves for the AL setting show how erroneous decisions can mislead the classifier, resulting in classifier accuracy below the baseline for two of the annotators, while the learning curves for these two annotators on the random samples show the same trend as for the other 4 annotators.

For *Konsequenz* (Figure 1, middle), the classifier trained on the AL samples yields results over the baseline after around 25 iterations, while in the random sampling setting it takes at least 100 iterations to beat the baseline. For *Motiv* (Figure 1, bottom row), again we observe far higher results in the AL setting. A possible explanation for why AL seems to work for *Ausgang, Motiv* and *Quelle* might be the higher IAA[4] (κ 0.825, 0.789, 0.768) as compared to the other target lemmas. This, however, does not explain the good results achieved on the AL samples for *Konsequenz,* for which IAA was quite low with κ 0.554.

Also startling is the fact that AL seems to work particularly well for one of the annotators (A_6, Figure 1) but not for others. Different possible explanations come to mind: (a) the accuracy of the annotations for this particular annotator, (b) the

[4] IAA was computed on the random samples, as the AL samples do not include the same instances.

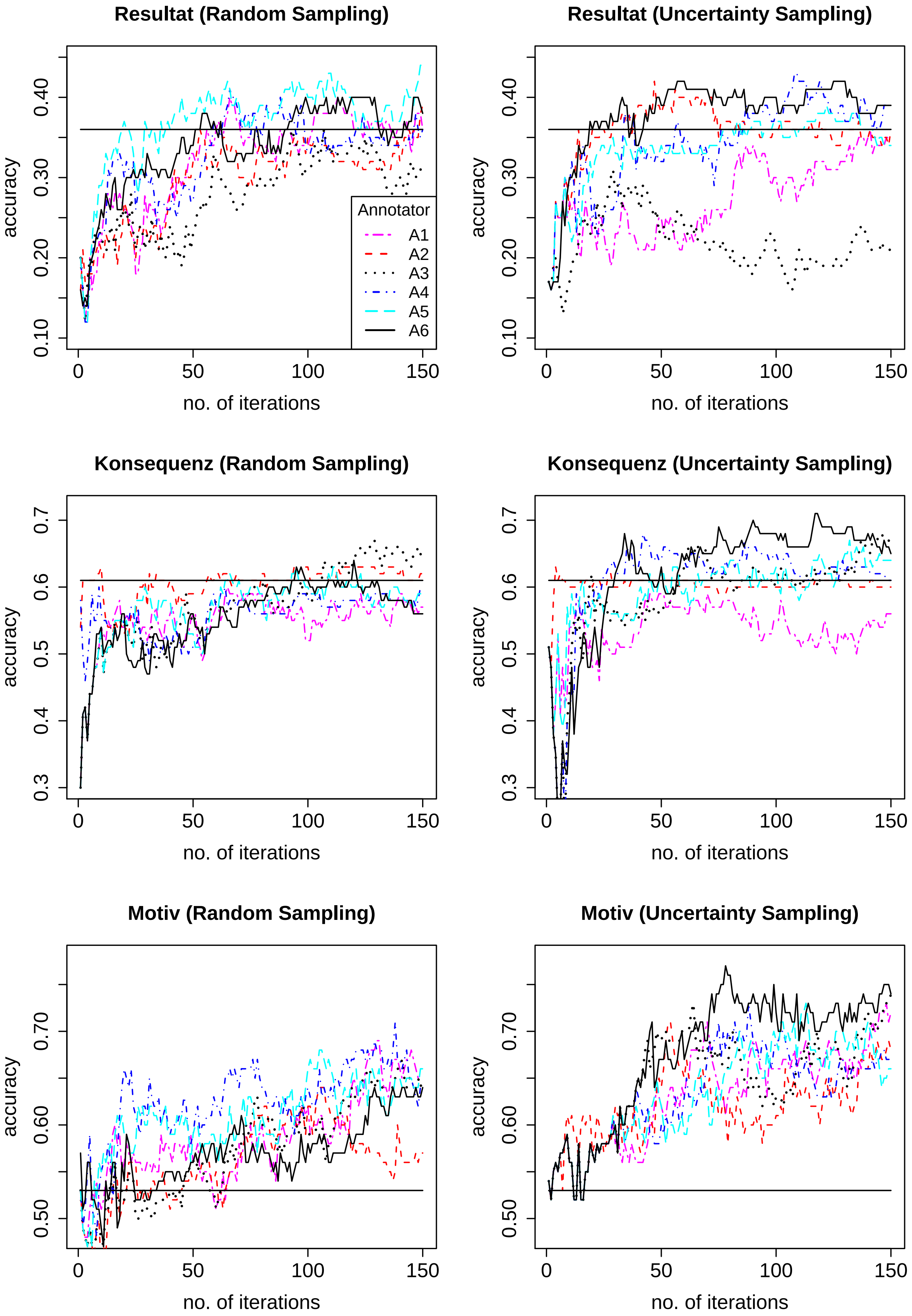

Figure 1: Active learning curves for Resultat, Konsequenz and Motiv (random sampling versus uncertainty sampling; the straight black line shows the majority baseline)

Konsequenz	A1	A2	A3	A4	A5	A6
human	0.80	0.72	0.89	0.73	0.89	0.76
maxent	0.60	0.63	0.67	0.60	0.63	0.64

Table 5: Acc. for human annotators against the adjudicated random samples and for the classifier

instances selected by the classifier based on the annotation decisions of the individual annotators, and (c) the distribution of frames in the annotated training sets for the different annotators.

To test (a) we evaluated the annotated random samples for *Konsequenz* for each annotator against the adjudicated gold standard. Results showed that there is no strong correlation between the accuracy of the human annotations and the performance of the classifier trained on these annotations. The annotator for whom AL worked best had a medium score of 0.76 only, while the annotator whose annotations were least helpful for the classifier showed a good accuracy of 0.80 against the gold standard.

Next we tested (b) the impact of the particular instances in the AL samples for the individual annotators on classifier performance. We took all instances in the AL data set from A_6, whose annotations gave the greatest boost to the classifier, removed the frame labels and gave them to the remaining annotators for re-annotation. Then we trained the classifier on each of the re-annotated samples and compared classifier performance. Results for 3 of the remaining annotators were in the same range or even higher than the ones for A_6 (Figure 2). For 2 annotators, however, results remained far below the baseline.

This again shows that the AL effect is not directly dependent on the accuracy of the individual annotators, but that particular instances are more informative for the classifier than others. Another crucial point is (c) the distribution of frames in the samples. In the annotated samples for A_1 and A_2 the majority frame for *Konsequenz* is Causation, while in the samples for the other annotators Response was more frequent. In our test set Response also is the most frequent frame, therefore it is not surprising that the classifiers trained on the samples of A_3 to A_6 show a higher performance. This means that high-quality annotations (identified by IAA) do not necessarily provide the information from which the classifier benefits most, and that in a realistic annotation task addressing the class imbalance problem (Zhu and Hovy, 2007) is crucial.

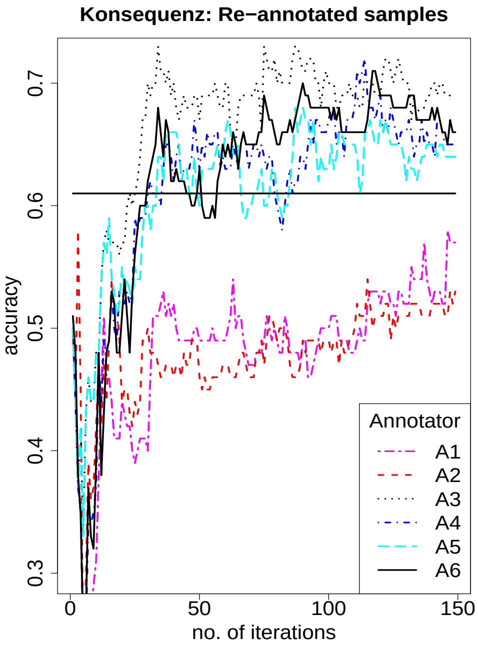

Figure 2: Re-annotated instances for Konsequenz (AL samples from annotator A_6)

6 Conclusions

We presented the first experiment applying AL in a real-world scenario by integrating the approach in an ongoing annotation project. The task and annotation environment pose specific challenges to the AL paradigm. We showed that annotation noise caused by biased annotators as well as erroneous annotations mislead the classifier and result in skewed data sets, and that for this particular task no time savings are to be expected when applied to a realistic scenario. Under certain conditions, however, classifier performance can improve over the random sampling baseline even on noisy data and thus yield higher accuracy in the active learning setting. Critical features which seem to influcence the outcome of AL are the amount of noise in the data as well as the distribution of frames in training- and test sets. Therefore, addressing the class imbalance problem is crucial for applying AL to a real annotation task.

Acknowledgments

This work was funded by the German Research Foundation DFG (grant PI 154/9-3 and the MMCI Cluster of Excellence).

References

Baldridge, Jason and Alexis Palmer. 2009. How well does active learning actually work?: Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of EMNLP 2009*.

Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of LREC-2006*.

Chan, Yee Seng and Hwee Tou Ng. 2007. Domain adaptation with active learning for word sense disambiguation. In *Proceedings of ACL-2007*.

Chen, Jinying, Andrew Schein, Lyle Ungar, and Martha Palmer. 2006. An empirical study of the behavior of active learning for word sense disambiguation. In *Proceedings of NAACL-2006*, New York, NY.

Cohn, David A., Zoubin Ghahramani, and Michael I. Jordan. 1995. Active learning with statistical models. In Tesauro, G., D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 705–712. The MIT Press.

Dang, Hoa Trang. 2004. *Investigations into the role of lexical semantics in word sense disambiguation*. PhD dissertation, University of Pennsylvania, Pennsylvania, PA.

Donmez, Pinar and Jaime G. Carbonell. 2008. Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of CIKM08*.

Erk, Katrin. 2005. Frame assignment as word sense disambiguation. In *Proceedings of the IWCS-6*.

Ertekin, Şeyda, Jian Huang, L'eon Bottou, and Lee Giles. 2007. Learning on the border: active learning in imbalanced data classification. In *Proceedings of CIKM '07*.

Hwa, Rebecca. 2004. Sample selection for statistical parsing. *Computational Linguistics*, 30(3):253–276.

Laws, Florian and Heinrich Schütze. 2008. Stopping criteria for active learning of named entity recognition. In *Proceedings of Coling 2008*.

Osborne, Miles and Jason Baldridge. 2004. Ensemble-based active learning for parse selection. In *Proceedings of HLT-NAACL 2004*.

Rehbein, Ines and Josef Ruppenhofer. 2010. Theres no data like more data? revisiting the impact of data size on a classification task. In *Proceedings of LREC-07, 2010*.

Reichart, Roi, Katrin Tomanek, Udo Hahn, and Ari Rappoport. 2008. Multi-task active learning for linguistic annotations. In *Proceedings of ACL-08: HLT*.

Ringger, Eric, Peter Mcclanahan, Robbie Haertel, George Busby, Marc Carmen, James Carroll, and Deryle Lonsdale. 2007. Active learning for part-of-speech tagging: Accelerating corpus annotation. In *Proceedings of ACL Linguistic Annotation Workshop*.

Tomanek, Katrin and Udo Hahn. 2009. Reducing class imbalance during active learning for named entity annotation. In *Proceedings of the 5th International Conference on Knowledge Capture*, Redondo Beach, CA.

Tomanek, Katrin, Joachim Wermter, and Udo Hahn. 2007. An approach to text corpus construction which cuts annotation costs and maintains corpus reusability of annotated data. In *Proceedings of EMNLP-CoNLL 2007*.

Zhu, Jingbo and Ed Hovy. 2007. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of EMNLP-CoNLL 2007*.

Zhu, Jingbo, Huizhen Wang, Tianshun Yao, and Benjamin K. Tsou. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of Coling 2008*.