Active learning for sense annotation

Héctor Martínez Alonso Barbara Plank Anders Johannsen Anders Søgaard
Njalsgade 140, Copenhagen (Denmark), University of Copenhagen
`alonso@hum.ku.dk, bplank@cst.dk, {ajohannsen, soegaard}@hum.ku.dk`

Abstract

This article describes a real (non-synthetic) active-learning experiment to obtain supersense annotations for Danish. We compare two instance selection strategies, namely lowest-prediction confidence (MAX), and sampling from the confidence distribution (SAMPLE). We evaluate their performance during the annotation process, across domains for the final resulting system, as well as against in-domain adjudicated data. The SAMPLE strategy yields competitive models that are more robust than the overly length-biased selection criterion of MAX.

1 Introduction

Most successful natural language processing (NLP) systems rely on a set of labeled training examples to induce models in a supervised manner. However, labeling instances to create a training set is time-consuming and expensive. One way to alleviate this problem is to resort to *active learning* (AL), where a learner chooses which instances—from a large pool of unlabeled data—to give to the human expert for annotation. After each annotation by the expert, the system retrains the learner, and the learner chooses a new instance to annotate.

There are many active learning strategies. The simplest and most widely used is *uncertainty sampling* (Lewis and Catlett, 1994), where the learner queries the instance it is most uncertain about (Scheffer and Wrobel, 2001; Culotta and McCallum, 2005). Instead, in *query-by-committee* an entire committee of models is used to select the examples with highest disagreement. At the same time most studies on active learning are actually *synthetic*, i.e. the human supervision was just emulated by holding out already labeled data.

In this study, we perform a *real* active learning experiment. Since speed plays a major role, we do not resort to an ensemble-based query-by-committee approach but use a single model for selection. We evaluate two selection strategies for a sequence tagging task, supersense tagging.

2 Datapoint-selection strategies

Given a pool of unlabeled data U, a datapoint-selection strategy chooses a new unlabeled item u_i to annotate. We evaluate two of such strategies. They both involve evaluating the informativeness of unlabeled instances.

The first strategy (MAX) is similar to the standard approach in uncertainty sampling, i.e. the active learning system selects datapoint whose classification confidence is the lowest. The second strategy (SAMPLE) attempts to make the selection criterion more flexible by sampling from the confidence score distribution.

The two strategies work as follows:

1. MAX: Predict on U and choose u_i that has the lowest prediction confidence p_i, where p_i is the posterior probability of the classifier for the item u_i.
2. SAMPLE: Predict on U and choose u_i sampling from the distribution of the inverse confidence scores for all the instances—making low-confidence items more likely to be sampled. We calculate the inverse confidence score as $-log(p_i)$.

We apply both datapoint-selection strategies on two different subcorpora sampled from the same unlabeled corpus (cf. Section 3). Each $(strategy, subcorpus)$ tuple yields a system setup for an individual annotator. Table 1 describes the setup of our four annotators.

3 Data collection

An AL setup requires some annotated data to use as training seed for the first model, and as evalua-

Annotator	Strategy	Subcorpus
A_{S_1}	SAMPLE	C_1
A_{S_2}	SAMPLE	C_2
A_{M_1}	MAX	C_1
A_{M_2}	MAX	C_2

Table 1: Annotators and their setup, namely their instance selection strategy and the unlabeled subcorpus.

Domain	$\overline{SL}$	Seed	Test
Blog	16.44		100
Chat	14.61		200
Forum	20.51		200
Magazine	19.45		200
Newswire	17.43	400	200
Parliament	31.21		200

Table 2: Super-sense tagging data sets

tion test bench. We use previously available Danish sense-annotated data (Martínez Alonso et al., 2015). This dataset is a subset of the the ClarinDK corpus (Asmussen and Halskov, 2012) and has been annotated by two annotators and later adjudicated by a third. Table 2 shows the different domains that make up the initial annotated data and how much is used for training seed and for testing. We choose a conventional scenario where the initial system is trained only on an usual kind of text (newswire) in order to later assess the system's improvement on out-of-domain data.

In addition to the labeled data used for training seed and for testing, we use two unlabeled 10K-sentence subcorpora. These two subcorpora (C_1 and C_2) are randomly sampled from the ClarinDK corpus in order to obtain sentences of the same type that make up the labeled data, but ensuring that the sentences in C_1 and C_2 do not overlap with any of the labeled sentences described in Table 2. All the sentences in C_1 and C_2 are between 5 and 50 words long, in order to limit the strong bias for selecting longer sentences in AL for sequence prediction.[1]

4 Model

The features used in the model are the following. For each word *w*, we calculate:

a) 2-TOKEN WINDOW of forms, lemmas and POS tags before and after *w*, including *w*.
b) 2-TOKEN WINDOW of most-frequent sense tags for *w*.
c) BAG OF WORDS of forms and lemmas at the *left* and *right* of *w*, marked for directionality so words at the left are different from words at the right.
d) MORPHOLOGY of *w*, whether it is all alphanumeric, capitalized, contains hyphens, and its 3-letter prefix and suffix.
e) BROWN CLUSTER estimated from U. We generate the 2,4,6,8,10 and 12-bits long prefix of the cluster bitstring of *w*.[2]

The system is trained using search-based classification (SEARN) (Daumé et al., 2009)[3] using default parameters and one pass over the data. We use one pass over the data, thereby using strict online learning.

5 Evaluation

The goal of an AL setup is to augment the set of training instances. In this section we evaluate the performance of the AL-generated annotations during the annotation process in form of learning curves (Section 5.1). In order to gauge the robustness of a system trained on data obtained from AL, we break the evaluation down by domain (Section 5.2). Finally, we compare a system trained exclusively on annotated and ajudicated newswire data with systems trained on a combination of training seed and AL data. We evaluate all systems on micro-averaged F1.

5.1 Learning curves

This section describes the life-cycle for the AL setup for the four annotators. The system does one pass over the data and then retrains after each item. We evaluate against the entire test data that comprises different domains (Section 5.2).

We delimit the learnability space of the task between the most-frequent-sense (MFS) baseline at the bottom and an estimate of an upper bound (UB). We approximate the UB by training a system on the seed plus all the four AL-annotated datasets. Note that the data for UB is four times the data at the end point of any learning curve.

[1] The data will be made available at clarin.dk under *Danish Supersense Annotation.*

[2] We use Liang's implementation https://github.com/percyliang/brown-cluster

[3] SEARN in Vowpal Wabbit https://github.com/JohnLangford/vowpal_wabbit/

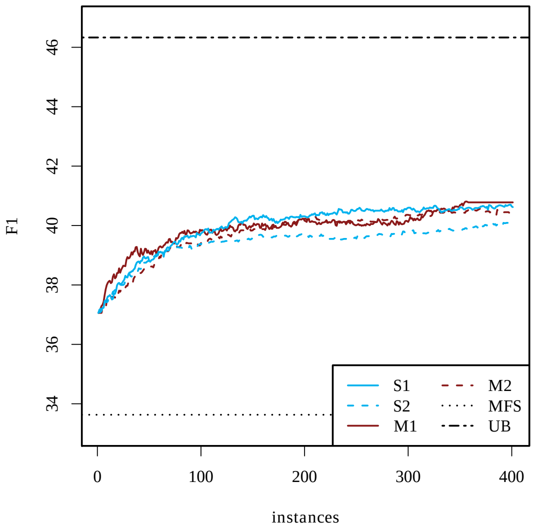

Figure 1: learning curves for the four annotators, delimited by the MFS baseline and the estimated upper bound (UB).

We observe that the MFS baseline (dotted line) is fairly low (33.63 F1). The system performance trained on seed starts at 37.06 F1. All learning curves show the same overall behavior with steeper learning for the first 150 instances. The differences are small, yet we can see that the SAMPLE approach surpasses the MAX strategy after 110 instances for one of the sub corpora. The most informative data during the initial iterations stems from annotator A_{M1} (MAX strategy), where we can observe the steepest increase. However, the MAX strategy results in *considerably* longer sentences (cf. $\overline{SL}$ in Table 4), thus is a major burden for the annotators compared to SAMPLE. In fact, A_{M1} could not finish the task in time, which is depicted by the straight line for the last 20 dots in the plot.

5.2 Performance across domains

The SAMPLE strategy turns out to be promising when evaluated on the entire test set, but only for one of the two subcorpora (Section 5.1). In this section, we look at the performance per domain.

Table 3 shows the domain-wise results for the last AL iteration of each annotator. We compare this to the most-frequent sense (MFS) baseline, as well to the original performance trained on only the seed data. The values for the annotators in Table 3 correspond with the last point of each learning curve in Figure 1, whereas the seed baseline is the common starting point (the intercept) for all learning curves.

We can see now that the seed baseline already beats MFS on all datasets except Chat, which is arguably very different from the newswire text that seed is sampled from. Overall, there is only a small difference in terms of performance between the two datapoint-selection strategies. The best strategy varies per domain.

			MAX		SAMPLE	
Dataset	MFS	Seed	$+A_{M_1}$	$+A_{M_2}$	$+A_{S_1}$	$+A_{S_2}$
Blog	25.6	37.2	40.5	**41.0**	*39.1*	39.4
Chat	36.1	*33.7*	39.8	*39.3*	40.1	**40.3**
Forum	31.1	33.4	**36.1**	35.9	*35.2*	35.5
Magazine	34.3	36.3	38.5	37.4	**38.6**	*36.8*
Newswire	31.5	37.2	**42.9**	40.5	41.2	*39.0*
Parliament	38.6	40.5	*45.0*	47.2	46.5	**47.5**
All	33.6	36.8	**40.8**	40.7	40.6	*40.1*

Table 3: Performance across domains for different training setups. The best system for each dataset is given in bold, and the worst system in italics.

Annotator	$\overline{SL}$	KL_{Seed-A}	KL_{A-Test}
A_{M_1}	43.36	0.026	0.027
A_{M_2}	42.65	0.019	0.027
A_{S_1}	26.16	0.035	0.054
A_{S_2}	26.24	0.017	0.022

Table 4: Descriptive statistics for the four generated datasets.

There are notable differences in terms of data characteristics between the resulting data samples. Table 4 shows descriptive statistics for the four resulting datasets, and how they relate to the seed training dataset and the overall test data. $\overline{SL}$ represents the average sentence length, KL_{Seed-A} is the Kullback-Leibler (KL) divergence from the seed dataset to each of the annotators', and KL_{A-Test} is the KL divergence from the annotators' datasets to the test data. We can see that there is a *big* difference in $\overline{SL}$ between the strategies. The system that use the SAMPLE strategy have a more variable sentence length but also much shorter sentences than MAX. It is still longer than the average sentence length in the unlabeled corpus, which is 21 tokens per sentence. This indicates that the selection strategy for SAMPLE is a compromise between the long-sentence bias of the MAX strategy and a purely random selection. Thus, we believe that the SAMPLE strategy is a viable alternative to the more common lowest-confidence strategy, because it can provide training data of competitive quality that is more varied and can provide more robust models.

5.3 Comparison with heldout data

In this last comparison, we gauge the effect of using adjudicated in-domain data instead of the same amount of AL-generated data. We remove the 200 newswire sentences from the test set and add them to the training seed. We train a system on the seed and these additional 200 adjudicated sentences (namely on all the 600 sentences of newswire data), and evaluate it across domains (all out-domain data). This system is compared to the result of the AL setup at the 200th iteration. The results in Table 5 show that even across domains it is more beneficial to have adjudicated in-domain data then the out-of-domain data annotated through active learning.

			MAX		SAMPLE	
Dataset	Seed	+Newswire	+A_{M_1}	+A_{M_2}	+A_{S_1}	+A_{S_2}
Blog	33.7	**39.9**	39.1	39.0	39.0	38.5
Chat	33.4	**41.9**	38.6	37.5	38.2	36.8
Forum	36.3	**39.2**	35.8	35.3	35.2	35.4
Magazine	40.5	**42.8**	39.7	39.4	40.1	39.4
Parliament	36.8	**48.2**	43.5	47.3	46.5	46.3
All	33.9	**43.2**	39.7	40.3	40.4	39.8

Table 5: Cross-domain performance against heldout newswire data

The causes for the better performance of the +Newswire system are twofold. First, there is less noise in the data because of the two-round process of annotation and adjudication; and second, the bias of this system's annotation is the same as in the evaluation data. Note that the 200-instance data point is past the 150-instance convergence point for the learning curves in Figure 1.

6 Related Work

The AL models considered here are very standard. We take a small seed of data points, train a sequential labeling, and iterate over an unlabeled pool of data, selecting the data points our labeler is least confident about. In the AL literature, the selected data points are often those close to a decision boundary or those most likely to decrease overall uncertainty. This obviously leads to biased sampling, which can sometimes be avoided using different techniques, e.g., by exploiting cluster structure in the data.

Generally, active learning for sequential labeling has received less attention than for classification (Settles and Craven, 2008; Marcheggiani and Artieres, 2014). Our experiments were simple, and several things can be done to improve results, i.e., by reducing sampling bias. In particular, several techniques have been introduced for improving out-of-domain performance using active learning. Rai et al. (2010) perform target-domain AL with a seed of source-domain data. Among other things, they propose to use the source and target unlabeled data to train a classifier to learn what target domain data points are similar to the source domain, in a way similar to Plank et al. (2014). For more work along these lines, see Chan and Ng (2007) and Xiao and Guo (2013).

7 Conclusions

The systems that use the MAX selection strategy have a strong bias for the longest possible sentence, resulting from the low probability values obtained when calculating the prediction confidence of very long sequences. With few exceptions (e.g. an 11-word sentence on the 5th iteration for A_{M_1}), the systems exhaust the maximum-length sentences, and proceed to choose the longest available, and so forth.

We do not take our individual annotator's bias into consideration, but we believe that such bias plays a minor role in the differences of performance between MAX and SAMPLE. For instance, A_{S_2} is the only annotator that was directly involved in the creation of the annotated seed and test data, but has arguably the worst-faring system on overall F1, that is, regardless of how *good* (i.e. how similar in bias to the test data) an annotator is, the selection strategy is the main factor for the improvement rate of the system during active learning. Note however that the data annotated by A_{S_2} does indeed have the lowest KL divergences to the seed and test data in Table 4.

Using SAMPLE lowers the annotator time per sentence because sentences do get shorter, even though the performance is initially lower until the 150-instance convergence point. We propose that with the same amount of time an annotator can annotate more, shorter sentences for the same amount of words and obtain more varied annotations that yield more robust models. Very long sentences do not bring a major advantage, and this justifies sampling when it is necessary to strike a compromise between annotation time and model robustness.

References

Jørg Asmussen and Jakob Halskov. 2012. The clarin dk reference corpus. In *Sprogteknologisk Workshop*.

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

Aron Culotta and Andrew McCallum. 2005. Reducing labeling effort for structured prediction tasks. In *AAAI*.

Hal Daumé, John Langford, and Daniel Marcu. 2009. Search-based structured prediction. *Machine learning*, 75(3):297–325.

David D Lewis and Jason Catlett. 1994. Heterogenous uncertainty sampling for supervised learning. In *ICML*.

Diego Marcheggiani and Thierry Artieres. 2014. An experimental comparison of active learning strategies for partially labeled sequences. In *EMNLP*.

Héctor Martínez Alonso, Anders Johannsen, Anders Søgaard, Sussi Olsen, Anna Braasch, Sanni Nimb, Nicolai Hartvig Sørensen, and Bolette Sandford Pedersen. 2015. Supersense tagging for danish. In *Nodalida*.

Barbara Plank, Anders Johannsen, and Anders Søgaard. 2014. Importance weighting and unsupervised domain adaptation of POS taggers: a negative result. In *EMNLP*.

Piyush Rai, Avishek Saha, Hal Daume, and Suresh Venkatasubramanian. 2010. Domain adaptation meets active learning. In *Workshop on Active Learning for NLP, NAACL*.

Tobias Scheffer and Stefan Wrobel. 2001. Active learning of partially hidden markov models. In *In Proceedings of the ECML/PKDD Workshop on Instance Selection*.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *EMNLP*.

Min Xiao and Yuhong Guo. 2013. Online active learning for cost sensitive domain adaptation. In *CoNLL*.