Experiments on Active Learning for Croatian Word Sense Disambiguation

Domagoj Alagić and **Jan Šnajder**
Text Analysis and Knowledge Engineering Lab
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Croatia
{domagoj.alagic,jan.snajder}@fer.hr

Abstract

Supervised word sense disambiguation (WSD) has been shown to achieve state-of-the-art results but at high annotation costs. Active learning can ameliorate that problem by allowing the model to dynamically choose the most informative word contexts for manual labeling. In this paper we investigate the use of active learning for Croatian WSD. We adopt a lexical sample approach and compile a corresponding sense-annotated dataset on which we evaluate our models. We carry out a detailed investigation of the different active learning setups, and show that labeling as few as 100 instances suffices to reach near-optimal performance.

1 Introduction

Word sense disambiguation (WSD) is the task of computationally determining the meaning of a word in its context (Navigli, 2009). WSD is considered one of the central tasks of natural language processing (NLP). A number of NLP applications can benefit from WSD, most notably machine translation (Carpuat and Wu, 2007), information retrieval (Stokoe et al., 2003), and information extraction (Markert and Nissim, 2007; Hassan et al., 2006; Ciaramita and Altun, 2006). At the same time, WSD is also considered a very difficult task; the difficulty arises from the fact that WSD relies on human knowledge and that it lends itself to different formalizations (e.g., the choice of a sense inventory) (Navigli, 2009).

The two main approaches to WSD are knowledge-based and supervised. Knowledge-based approaches rely on lexical knowledge bases such as WordNet. The drawback of knowledge-based approaches is that the construction of large-scale lexical resources requires a tremendous effort, rendering such approaches particularly cost-ineffective for smaller languages. On the other hand, supervised approaches do not rely on lexical resources and generally outperform knowledge-based approaches (Palmer et al., 2001; Snyder and Palmer, 2004; Pradhan et al., 2007). However, supervised methods instead require a large amount of hand-annotated data, which is also extremely expensive and time-consuming to obtain. Interestingly enough, Ng (1997) estimates that a wide coverage WSD system for English would require a sense-tagged corpus of 3200 words with 1000 instances per word. Assuming human throughput of one instance per minute (Edmonds, 2000), this amounts to an immense effort of 27 man-years.

One way of addressing the lack of manually sense-tagged data is to rely on semi-supervised learning (Abney, 2007), which, along with a smaller set of labeled data, also makes use of a typically much larger set of unlabeled data. A related technique is that of *active learning* (Olsson, 2009; Settles, 2010). However, what differentiates active learning from ordinary semi-supervised learning is that the former requires subsequent manual labeling. The underlying idea is to minimize the annotation effort by dynamically selecting the most informative unlabeled instances, i.e., the most informative contexts of a polysemous word to be manually labeled.

In this paper we address the WSD task for Croatian using active learning (AL). Croatian is an under-resourced language, lacking large-scale lexical resources and sense-annotated corpora. Our ultimate goal is to develop a cost-effective WSD system with a reasonable coverage for the most frequent Croatian words. As a first step towards that goal, in this paper we present a preliminary, small-scale but thorough empirical study using different AL setups. We adopt the lexical sample evaluation setup and evaluate our models on a chosen set of polysemous words. The contribution of our work is two-fold.

Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing, pages 49–58,
Hissar, Bulgaria, 10–11 September 2015.

First, we present a small sense-annotated dataset – the first such dataset for Croatian – which we also make freely available for research purposes. Secondly, we investigate in detail the performance of various AL models on this dataset and derive preliminary findings and recommendations. Although our focus is on Croatian, we believe our results generalize to other typologically similar (in particular Slavic) languages.

The rest of the paper is organized as follows. In the next section, we give a brief overview of AL-based WSD. In Section 3, we describe the manually sense-annotated dataset for Croatian. In Section 4, we describe the AL-based WSD models, while in Section 5 we present and discuss the experimental results. Lastly, Section 6 concludes the paper and outlines future work.

2 Related Work

WSD is a long-standing problem in NLP. A number of semi-supervised WSD methods have been proposed in the literature, including the use of external sources for the generation of sense-tagged data (McCarthy et al., 2004), 2004), use of bilingual corpora (Li and Li, 2004), label propagation (Niu et al., 2005), and bootstrapping (Mihalcea, 2004; Park et al., 2000).

Focusing on AL approaches to WSD, one of the first attempts is that of Chklovski and Mihalcea (2002). Their Open Mind World Expert system collected sense-annotated data over the web, which were later used for the Senseval-3 English lexical sample task (Mihalcea et al., 2004). The system employs the so-called committee-based sampling: the instances to be labeled are selected based on the disagreement between the labels assigned by two different classifiers.

Chen et al. (2006) experiment with WSD for five frequent English verbs. Unlike Chklovski and Mihalcea, they use uncertainty-based sampling coupled with a maximum entropy learner, and a rich set of topical, collocational, syntactic, and semantic features. Their results show that, given a target accuracy level, AL can reduce the number of training instances by half when compared to labeling randomly selected instances. Their analysis also reveals that careful feature design and generation is necessary to fully leverage the AL potential.

Additionally, a number of studies focus on issues specific to AL for WSD. Zhu and Hovy (2007) consider the class imbalance problem, which is typical for WSD due to skewness in sense distribution. They analyze the effect of resampling techniques and show that bootstrap-based oversampling of underrepresented senses improves classifier performance. Another important issue of AL is the stopping condition. Zhu and Hovy (2007) propose a stopping criterion based on the classifier confidence, Wang et al. (2008) propose a minimum expected error strategy, while Zhu et al. (2008a) propose classifier-change as a stopping criterion. Finally, Zhu et al. (2008b) propose sampling methods for generating a representative initial training set, as well as selective sampling method for alleviating the problem of outliers.

All of the above cited work addresses WSD for English, whereas our work focuses on Croatian. Similar to Chen et al. (2006), we use uncertainty-based sampling but combine it with an SVM model. In contrast to Chen et al. (2006), we opt for simple, readily available features derived from co-occurrences. We study three sampling methods in this work, but leave the issues of stopping criterion and class imbalance for future work.

Croatian is a Slavic language, and WSD for Slavic languages seems not to have received much attention so far. Notable exceptions are (Baś et al., 2008; Broda and Piasecki, 2009) for Polish and (Lyashevskaya et al., 2011) for Russian. WSD for Bulgarian, Czech, Serbian, and Slovene has been considered in a cross-lingual setup by Tufiş et al. (2004) and Ide et al. (2002). Bakarić et al. (2007) analyze the discriminative strength of co-occurring words for WSD of Croatian nouns. Additionally, Karan et al. (2012) consider a problem dual to WSD, namely synonymy detection. To the best of our knowledge, our work is the first reported work on active learning for WSD for a Slavic language.

3 Dataset

In this work we adopt the lexical sample style evaluation, i.e., we select a set of words and sample sentences from a corpus containing these words. We next describe how we compiled and sense-annotated the sample.

3.1 Corpus and Preprocessing

To compile a sense-annotated dataset for our experiments, we sample from a Croatian web corpus hrWaC[1] (Ljubešić and Klubička, 2014), containing 1.9M tokens, annotated with lemma, morphosyntax and dependency syntax tags.

For the sense inventory, we initially adopted the Croatian wordnet (CroWN) compiled by Raffaelli et al. (2008). Although of a limited coverage (10k synsets, compared to 200k synsets of Princeton WordNet), CroWN was a good starting point for word selection and sense definition for this task.

To keep the annotation effort manageable, similarly to (Chen et al., 2006), we decided to limit ourselves to six words: two nouns, two verbs, and two adjectives. We selected these by first compiling a list of polysemous words from CroWN that occur at least 1000 times in hrWaC. We then decided to discard words with more than three senses as our preliminary analysis revealed that CroWN senses of such words are potentially very difficult to differentiate. The problem of sense granularity of wordnets is a well-known issue (Edmonds and Kilgarriff, 2002), and in this study we wanted to avoid the problem by choosing words with as distinct senses as possible.[2] Research on sense granularity in the context of AL is warranted but is beyond the scope of this paper.

The final list of words is as follows: $okvir_N$ (frame), $vatra_N$ (fire), $brusiti_V$ (to rasp), $odlikovati_V$ (to award), lak_A (easy), and $prljav_A$ (dirty). For each of these words, we sampled 500 sentences from hrWaC, yielding a total of 3000 word instances. Note that 500 instances per word is well above the $75+15\cdot n$ instances recommended by Edmonds and Cotton (2001), where n is the number of senses of the word.

3.2 Sense Annotation

To construct the sense-annotated dataset, we asked 10 annotators to label the senses of the selected words in sampled sentences. Each annotator was given 600 sentences to annotate, with 100 instances of each of the six words. To obtain a more reliable annotation, each instance was double-annotated, and we ensured that there is a uniform distribution across the annotator pairings.

For each word instance, the annotators were given a list of possible word senses (two or three) and an additional "none of the above" (NOTA) option. They were instructed to select a single sense, unless there is no adequate sense listed or the instance is erroneous (incorrect lemmatization or a spelling error). For each sense, we provided a gloss line and usage examples from CroWN.

The annotation guidelines were rather straightforward. In cases when more than a single sense apply, the annotators were asked to choose the one they deem more appropriate. The only issue that we felt deserved additional elaboration was the treatment of polysemous words in semantically opaque contexts (idioms and metaphors). In such contexts, the annotators were asked to choose the literate sense of a word, rather than to consider the idiomatic or metaphoric reading. For example, in sentence *Istarska kuhinja je dijamant koji treba brusiti* (*Istrian cuisine is a diamond that needs to be cut*), the verb *brusiti* (*to cut* in this example) is used in its literate sense (*to rasp*), although the whole phrase *brusiti dijamant* is used in a metaphorical sense, which in this case happens to be somewhat related to the *to hone* sense of *brusiti*.[3]

The total effort for annotating 6000 word instances (including double annotations) was 36 man-hours, i.e., a throughput of 22 seconds per word instance. We note that this is considerably lower than the one-minute-per-instance estimate of Edmonds (2000). One of the possible reasons for this difference might be the biased word selection process, which probably resulted in somewhat easier disambiguation tasks.

3.3 Inter-Annotator Agreement

We use Cohen's kappa to measure the inter-annotator agreement (IAA). We calculate the agreement for each word separately by averaging the agreements for each annotator pair that labeled that word. The per-word IAA is shown in Table 1. The average IAA across the six words is 0.761, which, according to Landis and Koch (1977) is considered a substantial agreement.

Two words that stand out in terms of IAA are *odlikovati* (high IAA) and *prljav* (low IAA). The former has two clearly distinguishable senses. The latter turned out to be problematic as the word is of-

[1] http://nlp.ffzg.hr/resources/corpora/hrwac/

[2] We are aware that selecting words with easily distinguishable senses results in a biased sample. However, we note that such a sample does not necessarily need to be *unrealistically* easy. One could argue that senses that are difficult to differentiate are not realistic to begin with, as they are not likely to be of practical interest in real-world NLP applications.

[3] The alternative strategy would be to exclude (ask the annotators to tag as NOTA) all instances with opaque contexts, under the justification that idioms and metaphors require a special treatment. We will investigate this strategy in future work.

Word	κ	Word	κ
okvir$_N$	0.795	*odlikovati*$_V$	**0.978**
vatra$_N$	0.704	*lak*$_A$	0.582
brusiti$_V$	0.816	*prljav*$_A$	0.690

Table 1: Cohen's κ for the six chosen words.

Word	Freq.	# Senses	Sense distr.	NOTA
okvir$_N$	141862	2	381 / 115	4
vatra$_N$	45943	3	244 / 106 / 141	9
brusiti$_V$	1514	3	205 / 262 / 27	7
odlikovati$_V$	15504	2	425 / 75	0
lak$_A$	15424	3	277 / 87 / 113	23
prljav$_A$	14245	2	228 / 187	85

Table 2: Statistics of the gold standard sample.

ten used as part of the idiomatic expression *prljavo rublje* (*dirty laundry*). According to our annotation guidelines, here *prljav* is used in its literal sense (*dirty*), as *dirty laundry* is an idiom (matters embarrassing if made public). Annotators often selected the other, "sordid" meaning of *prljavi*, probably because they felt it is more related to the meaning of the idiom. Another source of disagreement are the named entities *Prljavo kazalište* (a rock band) and *Prljavi Harry* (the movie *Dirty Harry*), in which the intended sense of *prljavo* is questionable.

3.4 Gold Standard Sample

The last step in data annotation was to manually resolve the disagreements and obtain a gold standard sample. While trying to resolve the disagreements, we noticed that a large number of them are systematic – most of the time, one of the two annotators chose the NOTA option. Upon closer inspection, we found that for the most of the six considered words the CroWN sense inventory was arguably incomplete. To overcome this problem, we decided to modify the CroWN sense inventory for the six considered words to get a reasonable sense coverage on our lexical sample. Using this revised sense inventory, we (the authors) resolved all the disagreements (a 6 man-hours effort). The statistics of the 3000-sentences gold standard sample is shown in Table 2. Sense inventory is given in Table 3. We make the dataset freely available.[4]

okvir (frame)	
#1	An environment to which the situation is related or whose influence it is exposed to.
#2	A structure that supports or contains something.
vatra (fire)	
#1	One of the four fundamental classical elements (along with water, air, and earth) according to Empedocles.
#2	The act of firing weapons or artillery at an enemy.
#3	A heat source for food preparation.
brusiti (to rasp)	
#1	Making something smooth using a file or a rasp.
#2	Gaining skill in something; taking quality, readiness, and specific knowledge and abilities to a high level.
#3	Increasing the level of eagerness/tension/excitement.
odlikovati (to award)	
#1	Having a certain characteristic, trait, feature.
#2	Giving something to someone, especially as a reward for an accomplishment.
lak (easy)	
#1	One that does not require a lot of effort to be carried out or understood.
#2	One that possesses a small physical mass.
#3	One that is not strong or intense.
prljav (dirty)	
#1	One that contains or produces stains or filth.
#2	One that is not morally pure.

Table 3: Sense inventory.

4 Models

4.1 Active Learning Setup

There are a number of different AL strategies; refer to Settles (2010) for a comprehensive overview. We employ the *pool-based strategy* (Lewis and Gale, 1994) using *uncertainty sampling*. This method uses a small set of labeled data L (the seed train set) and a large pool of unlabeled data U. The classifier is first trained on set L. After that, P (the pool size) instances are randomly sampled from U and the classifier is used to predict their labels. Next, from this set at most G (train growth size) instances are selected for which the classifier is the least confident about and an oracle (e.g., a human expert) is queried for their labels. Finally, the newly-labeled instances are added to the training set L and the process is repeated until a stopping criterion is met. The active learning loop is shown in Algorithm 1.

The motivation for the use of a pool is to reduce the computational cost associated with sense label prediction on the entire set of unlabeled instances

[4] http://takelab.fer.hr/cro6wsd

Algorithm 1: Active learning loop

```
$L$ : initial training set
$U$ : pool of unlabeled instances
$P$ : pool sample size
$G$ : train growth size
$f$ : classifier
while stopping criteria not satisfied do
    $f \leftarrow$ train($f$, $L$);
    $R \leftarrow$ randomSample($U$, $P$)
    predictions $\leftarrow$ predict($f$, $R$)
    $R \leftarrow$ sortByUncertainty($R$, predictions)
    $S \leftarrow$ selectTop($R$, $G$)
    $S \leftarrow$ oracleLabel($S$)
    $L \leftarrow L \cup S$
    $U \leftarrow U \setminus S$
end
```

U. In our experiments, U is relatively small, thus we decide to use the complete set U as the pool, $P = |U|$. This eliminates one source of randomness and allows us to focus on other, in our view, more important AL parameters.

Our experiments are focused on different uncertainty sampling methods. We therefore simulate a perfect oracle by providing the labels from the gold standard sample for each query. Furthermore, we ignore the stopping criterion issue and run the AL algorithm until the complete training set is utilized.

We consider three uncertainty sampling methods, i.e., methods for evaluating the informativeness of an unlabeled instance, as outlined below.

Least confident (LC). Trivially, the most informative instance is the one for which the prediction is the least confident:

$$x^*_{\text{LC}} = \operatorname*{argmax}_x \big(1 - P_\theta(\hat{y}|x)\big) \qquad (1)$$

where $\hat{y}$ stands for the class label with the highest posterior probability under the model θ.

Minimum margin (MM). An instance for which the difference between the posterior probabilities of two most probable class labels is maximal bears the most information:

$$x^*_{\text{MM}} = \operatorname*{argmin}_x \big(P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x)\big) \qquad (2)$$

where $\hat{y}_1$ and $\hat{y}_2$ are the first and second most probable class labels under the model θ.

Maximum entropy (ME). Selects an instance whose vector of posterior class label probabilities has the maximum entropy:

$$x^*_{\text{ME}} = \operatorname*{argmax}_x \big(- \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x)\big) \qquad (3)$$

4.2 Classifier and Features

As the core classifier in AL experiments, we use a linear Support Vector Machine (SVM) implemented in LIBSVM (Chang and Lin, 2011) library. To turn SVM confidence scores into probabilities over classes, we use the method proposed by Platt (1999), also implemented in the same library. Multiclass classification is handled using the *one-vs-one* scheme.

We opt for a simple model with readily available features. The simplest features are word-based context representations: given a sentence in which a polysemous word occurs, we compute its context vector by considering the words it co-occurs with in the sentence. We consider two context representations. First is a simple binary bag-of-words vector (BoW). In our case, the average dimension of a BoW vector is approximately 7000.

The second representation we use is the recently proposed skip-gram model, a neural word embedding method of Mikolov et al. (2013), which has shown to be useful on a series of NLP tasks. To obtain a context vector, we simply add up the skip-gram vectors of all the context words. The advantage of skip-gram representation over BoW is that it generates compact, continuous, and distributed vectors representations such that semantically related words tend to have similar vectors. This not only results in more effective context representations, but also allows for a better generalization, as context vectors of words unseen during training will be similar to vectors of semantically related context words used for training. We build the vectors from hrWaC using the `word2vec`[5] tool. We use 300 dimensions, negative sampling parameter set to 5, minimum frequency set to 100, and no hierarchical softmax.

5 Experimental Results

In this section we describe the AL experiments on our lexical sample dataset. We randomly split the dataset into a training and a test set: for each of the six words, we use 400 instances for training and 100 for testing.

5.1 Supervised Baseline

We compare our AL-based models against their fully supervised counterparts as baselines, i.e., linear SVM classifiers with either BoW or skip-gram context representations, denoted SVM-BoW and

[5] `https://code.google.com/p/word2vec/`

Word	MFS	SVM-BoW	SVM-SG
okvir$_N$	0.53	**0.92**	0.89
vatra$_N$	0.49	**0.91**	0.88
brusiti$_V$	0.53	0.85	**0.86**
odlikovati$_V$	0.85	**0.97**	**0.97**
lak$_A$	0.55	0.80	**0.81**
prljav$_A$	0.46	0.82	**0.88**
Average:	0.57	**0.88**	**0.88**

Table 4: Supervised models accuracy.

SVM-SG, respectively. In addition, we use the most frequent sense (MFS) as a baseline for the supervised models. Note that MFS has been generally proven to be a very strong baseline for WSD. We optimized the SVM regularization parameter C using 5-fold cross-validation on the training set.

Table 4 shows the results on the test set. Overall, the SVM models perform comparably well and outperform the MFS baseline by a wide margin. The models perform best on *odlikovati*, which was also the word with the highest IAA score (cf. section 3.2). The MFS baseline also performs quite well on this word due to its skewed sense distribution.

5.2 Active Learning Experiments

For AL experiments we use the same train-test split as before. The difference is that, for each word, the initial training set L is a randomly chosen subset of the full training set. In what follows, to obtain robust performance estimates, we run 50 trials of each experiment, each time random sampling anew the set L, and then averaging the results.

AL is governed by a number of parameters: the choice of the sampling method, train growth size G, and the size of the initial training set L. To more clearly show the effectiveness of AL, we set G to 1 and the size of the initial training set to 20, but elaborate on this choice later.

For the C parameter we use the same value as above, i.e., the value optimized using cross-validation on the entire training set. Arguably, this is not a realistic AL setup, as in practice the entire training is not labeled up front. In this work, however, we decided to simplify the setup as we observed that on our dataset the optimal C value is rather stable regardless of the training set size.

Learning curves. The purpose of AL is to reduce the labeling effort, i.e., to achieve a satisfactory level of accuracy with a smaller number of training instances. To analyze the effectiveness of AL WSD on our lexical sample, we compute the learning curves for SVM-BoW and SVM-SG and the three considered uncertainty sampling methods. The baselines are the learning curves obtained using random sampling (RAND). Fig. 1 shows the learning curves and the standard deviation bands.

The first thing we observe is that all uncertainty sampling methods outperform RAND. For example, when the training set reaches 100 instances, AL with uncertainty sampling outperforms RAND by $\sim$2% of accuracy for both SVM-BoW and SVM-SG models. In our view, this performance gain justifies the use of AL WSD on our dataset.

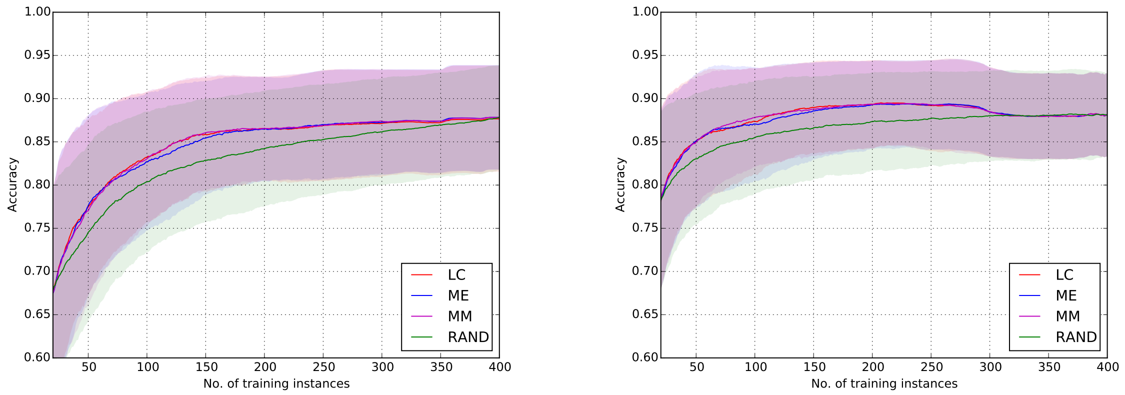
The second thing we observe is that the three uncertainty sampling methods generally perform comparably. However, the least confident (LC) and maximum margin (MM) methods slightly outperform the maximum entropy (ME) method in the 100–150 instances range.

The last thing we observe is that, with uncertainty sampling, labeling as few as 100 out of 400 training instances already gives $\sim$0.94% of maximum accuracy for SVM-BoW, while random sampling requires a training set of twice that size. Moreover, labeling 150 instances gives almost maximum accuracy for SVM-BoW. For SVM-SG, the effect of uncertainty sampling is even more pronounced – with 100 instances we already reach performance equivalent to that on the full training set. We conclude that AL WSD with SVM-SG reduces the number of training instances to 100 without any drop in performance.

Taking into account the previous observations, we decided to use the SVM-SG model and MM uncertainty sampling in subsequent experiments.

Parameter analysis. To investigate the impact of the initial training set size L and the train growth size G, we run a grid search with $L \in \{20, 50, 100\}$ and $G \in \{1, 5, 10\}$. For each pair of parameter values, we carry out 50 AL runs per word, each time using a random sample of size L as the initial training set. We thus obtain a total of 300 runs per parameter pair, which we average to produce corresponding learning curves. We compare the AL WSD performance in terms of the Area Under Learning Curve (ALC), which we define as a sum of classifier accuracy scores across the iterations of the AL algorithm, normalized by the number of iterations.

Table 5 shows the ALC scores for different parameter combinations. Expectedly, the larger the initial training set L, the more information is avail-



(a)

(b)

Figure 1: Learning curves for (a) SVM-BoW and (b) SVM-SG.

	G		
$\lvert L \rvert$	1	5	10
20	0.8794	0.8772	0.8760
50	0.8824	0.8819	0.8810
100	**0.8843**	0.8836	0.8833

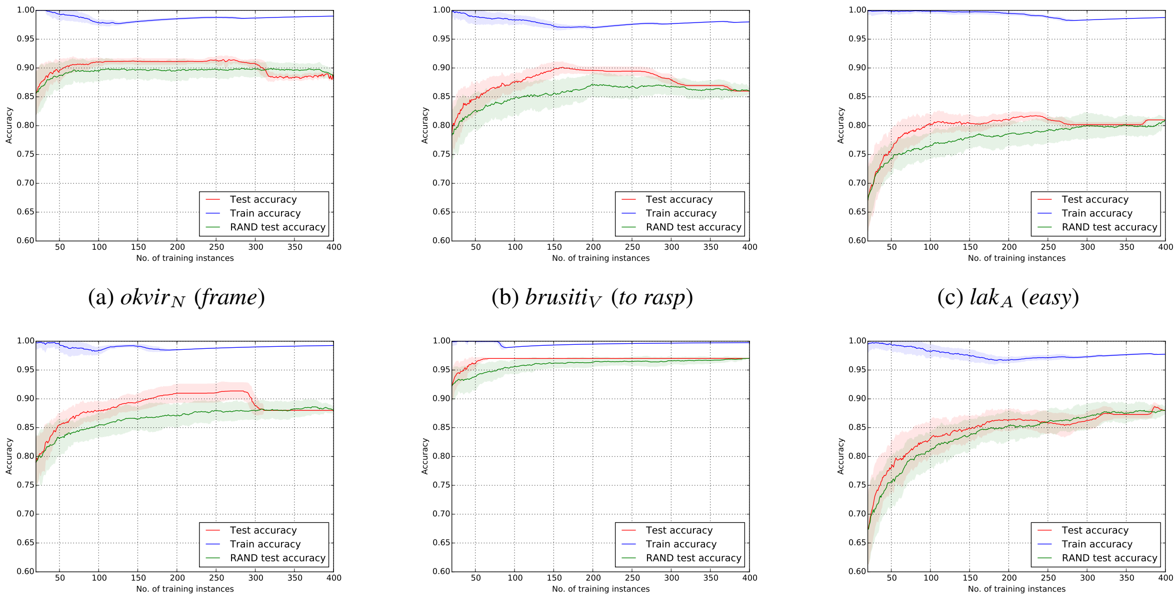
Table 5: ALC scores across parameters for SVM-SG with MM sampling.

able to the learning algorithm up front. At the same time, using a train growth size *G* of one yields better models, as they are able to make more confident predictions on yet unlabeled instances in each iteration of the AL algorithm. Nonetheless, we observe that in our case these two AL parameters do not considerably affect the model performance.

Per word analysis. In the previous analyses we looked at learning curves averaged over the six words in our dataset. For a more detailed analysis, we turn to the learning curves of the individual words, shown in Fig. 2. We plot both the accuracy on the training set and the test set using the MM sampling method, as well as the RAND accuracy on the test set. Note that a large gap between the curves on the training set and test set indicates model overfitting.

The plots reveal that MM outperforms the RAND baseline for all six words. Moreover, the gain is most prominent for *vatra*, *lak*, and *brusiti*. On *odlikovati* the full maximum accuracy can already be reached with as few as 60 training instances. In contrast, the word *prljav* is a problematic one: the learning curve does not seem to get saturated even after 400 instances. This is probably due to the many NOTA labels for that words. The train-test curve gap is the largest for *lak*, suggesting that the model overfits the most on that particular word. We hypothesize that, for some reason, the instances of this word are more noisy than instances of other words. Because disagreements in our dataset have been manually resolved, we think that latent variables are a more likely explanation for the noise than mislabelings. In other words, we believe that for some reason skip-gram contexts are less informative of the senses of the word *lak* than of the other words.

Another interesting observation is that for some words the accuracy rises above that of a model trained on the entire training set of 400 instances, after which it drops and eventually the two accuracy curves converge. This effect is most prominent for *vatra* and *brusiti*, and somewhat less for *okvir* and *lak*. A similar effect has been observed by Chen et al. (2006) on some English verbs, suggesting that the effect can be traced down to model starting to overfit at some point. We think that this hypothesis is plausible, as it is also confirmed by the fact that we observe no drop in the training error. Moreover, we hypothesize that the drop in accuracy is due to the sampling of a sequence of noisy examples from the training set. By the same token as before, we tend to exclude mislabelings as the cause of the noise, but rather attribute the noise to non-informative contexts. The existence of such "bad examples" was hypothesized by Chen et al. (2006), who suggest that that adequate feature selection could solve the problem. We leave a detailed investigation for future work.



(a) *okvir*$_N$ (*frame*)

(b) *brusiti*$_V$ (*to rasp*)

(c) *lak*$_A$ (*easy*)

(d) *vatra*$_N$ (*fire*)

(e) *odlikovati*$_V$ (*to award*)

(f) *prljav*$_A$ (*dirty*)

Figure 2: Learning curves for words from the lexical sample.

6 Conclusion

We have explored the use of active learning (AL) in Croatian word sense disambiguation (WSD). We manually compiled a sense-annotated dataset of six polysemous words. On this dataset, we have shown that by using uncertainty-based sampling we can reach a 99% of accuracy of a fully supervised model at the cost of annotating only 100 instances. On some words, the AL WSD even outperforms a fully supervised model.

Our main priority for future work is to extend our lexical sample. Having a more representative dataset at our disposal, we plan to study how AL WSD performance relates to the linguistic properties of polysemous words, and how these can be exploited to improve the sampling of instances. We also plan to investigate the issues of class imbalance, stopping criteria, and other uncertainty sampling methods.

Having in mind our ultimate goal of creating a cost-effective WSD for Croatian, another interesting direction for future work is to study AL WSD in a crowdsourcing (noisy multi-annotator) environment.

Acknowledgments

We are grateful to the ten annotators for their hard work, as well as to the three anonymous reviewers for their insightful comments and suggestions.

References

Steven Abney. 2007. *Semisupervised learning for computational linguistics*. CRC Press.

Nikola Bakarić, Jasmina Njavro, and Nikola Ljubešić. 2007. What makes sense? Searching for strong WSD predictors in Croatian. In *INFuture2007: Digital Information and Heritage*, pages 321–326.

Dominik Baś, Bartosz Broda, and Maciej Piasecki. 2008. Towards word sense disambiguation of Polish. In *Proceedings of IMCSIT*, pages 73–78, Wisla, Poland.

Bartosz Broda and Maciej Piasecki. 2009. Semi-supervised word sense disambiguation based on weakly controlled sense induction. In *Proceedings of IMCSIT*, pages 17–24, Mragowo, Poland.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of EMNLP-CoNLL*, volume 7, pages 61–72, Prague, Czech Republic.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Jinying Chen, Andrew Schein, Lyle Ungar, and Martha Palmer. 2006. An empirical study of the behavior of active learning for word sense disambiguation. In *Proceedings of HLT-NAACL*, pages 120–127, New York, USA.

Timothy Chklovski and Rada Mihalcea. 2002. Building a sense tagged corpus with Open Mind Word

Expert. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, volume 8, pages 116–122, Philadelphia, USA.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of EMNLP*, pages 594–602, Sydney, Australia.

Philip Edmonds and Scott Cotton. 2001. Senseval-2: overview. In *Proceedings of SensEval-2*, pages 1–5, Toulouse, France.

Philip Edmonds and Adam Kilgarriff. 2002. Introduction to the special issue on evaluating word sense disambiguation systems. *Natural Language Engineering*, 8(04):279–291.

Philip Edmonds. 2000. Designing a task for Senseval-2.

Hany Hassan, Ahmed Hassan, and Sara Noeman. 2006. Graph based semi-supervised approach for information extraction. In *Proceedings of TextGraphs-1*, pages 9–16, New York, USA.

Nancy Ide, Tomaž Erjavec, and Dan Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, volume 8, pages 61–66, Philadelphia, USA.

Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Distributional semantics approach to detecting synonyms in Croatian language. *Information Society*, pages 111–116.

J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of ACM SIGIR*, pages 3–12, Dublin, Ireland.

Hang Li and Cong Li. 2004. Word translation disambiguation using bilingual bootstrapping. *Computational Linguistics*, 30(1):1–22.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of WaC*, pages 29–35, Gothenburg, Sweden.

Olga Lyashevskaya, Olga Mitrofanova, Maria Grachkova, Sergey Romanov, Anastasia Shimorina, and Alexandra Shurygina. 2011. Automatic word sense disambiguation and construction identification based on corpus multilevel annotation. In *Text, Speech and Dialogue*, pages 80–90.

Katja Markert and Malvina Nissim. 2007. Semeval-2007 task 08: Metonymy resolution at semeval-2007. In *Proceedings of SemEval-2007*, pages 36–41, Prague, Czech Republic.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of ACL*, pages 279–286, Barcelona, Spain.

Rada Mihalcea, Timothy Anatolievich Chklovski, and Adam Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of SensEval-3*.

Rada Mihalcea. 2004. Co-training and self-training for word sense disambiguation. In *Proceedings of CoNLL*, pages 33–40, Boston, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119, Nevada, USA.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

Hwee Tou Ng. 1997. Getting serious about word sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, pages 1–7, Washington, D.C., USA.

Zheng-Yu Niu, Dong-Hong Ji, and Chew Lim Tan. 2005. Word sense disambiguation using label propagation based semi-supervised learning. In *Proceedings of ACL*, pages 395–402, Michigan, USA.

Fredrik Olsson. 2009. A literature survey of active machine learning in the context of natural language processing. Technical report, Swedish Institute of Computer Science, Kista, Sweden.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of SenseEval-2*, pages 21–24, Toulouse, France.

Seong-Bae Park, Byoung-Tak Zhang, and Yung Taek Kim. 2000. Word sense disambiguation by learning from unlabeled data. In *Proceedings of ACL*, pages 547–554, Hong Kong, China.

John C Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74.

Sameer S Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task 17: English lexical sample, SRL and all words. In *Proceedings of SemEval-2007*, pages 87–92, Prague, Czech Republic.

Ida Raffaelli, Marko Tadić, Božo Bekavac, and Željko Agić. 2008. Building Croatian wordnet. In *Proceedings of GWC*, pages 349–360, Szeged, Hungary.

Burr Settles. 2010. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of Senseval-3*, pages 41–43, Barcelona, Spain.

Christopher Stokoe, Michael P Oakes, and John Tait. 2003. Word sense disambiguation in information retrieval revisited. In *Proceedings of ACM SIGIR*, pages 159–166, Toronto, Canada.

Dan Tufiş, Radu Ion, and Nancy Ide. 2004. Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *Proceedings of COLING*, pages 1312–1318, Geneva, Switzerland.

Huizhen Wang, Jingbo Zhu, and Eduard Hovy. 2008. Learning a stopping criterion for active learning for word sense disambiguation and text classification. In *Proceedings of IJCNLP*, pages 366–372, Hyderabad, India.

Jingbo Zhu and Eduard H Hovy. 2007. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of EMNLP-CoNLL*, volume 7, pages 783–790, Prague, Czech Republic.

Jingbo Zhu, Huizhen Wang, and Eduard Hovy. 2008a. Multi-criteria-based strategy to stop active learning for data annotation. In *Proceedings of COLING*, volume 1, pages 1129–1136, Manchester, UK.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. 2008b. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of COLING*, volume 1, pages 1137–1144, Manchester, UK.