Cro36WSD: A Lexical Sample for Croatian Word Sense Disambiguation

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

Abstract

We introduce Cro36WSD, a freely-available medium-sized lexical sample for Croatian word sense disambiguation (WSD). Cro36WSD comprises 36 words: 12 adjectives, 12 nouns, and 12 verbs, balanced across both frequency bands and polysemy levels. We adopt the multi-label annotation scheme in the hope of lessening the drawbacks of discrete sense inventories and obtaining more realistic annotations from human experts. Sense-annotated data is collected through multiple annotation rounds to ensure high-quality annotations: with a 115 person-hours effort we reached an inter-annotator agreement score of 0.877. We analyze the obtained data and perform a correlation analysis between several relevant variables, including word frequency, number of senses, sense distribution skewness, average annotation time, and the observed inter-annotator agreement (IAA). Using the obtained data, we compile multi- and single-labeled dataset variants using different label aggregation schemes. Finally, we evaluate three different baseline WSD models on both dataset variants and report on the insights gained. We make both dataset variants freely available.

Keywords: word sense disambiguation, lexical sample, multi-label annotation, Croatian language

1. Introduction

Word sense disambiguation (WSD), a task of computationally determining the meaning of a word in its context (Navigli, 2009), is one of the longest-standing and most crucial tasks of natural language processing (NLP). Knowing the right sense of a word can be beneficial in various NLP applications, such as machine translation (Carpuat and Wu, 2007), information retrieval (Stokoe et al., 2003), and information extraction (Ciaramita and Altun, 2006).

An indispensable ingredient in the development and testing of WSD systems are sense-annotated corpora. Unfortunately, such corpora are extremely costly to produce, mainly because a sufficient number of contexts has to be manually labeled for each polysemous word. Moreover, most work on WSD has focused on English and has relied on WSD datasets built as part of the SemEval (Senseval) evaluation exercises (Navigli et al., 2007; Agirre et al., 2009; Manandhar et al., 2010; Moro and Navigli, 2015). Consequently, WSD datasets for languages other than English are comparably rare.

In this paper, we present Cro36WSD – a medium-sized lexical sample dataset for Croatian WSD, which extends our Cro6WSD dataset introduced in (Alagić and Šnajder, 2015). Extensions include a larger number of words in the lexical sample, adoption of the multi-label annotation scheme, and the sense inventory more apt to the task. The latter two extensions are motivated by the often-discussed inadequateness of discrete sense inventories (Erk and McCarthy, 2009) and their granularities (Edmonds and Kilgarriff, 2002; Snyder and Palmer, 2004). We construct two different variants of the dataset according to different label aggregation schemes. We also carry out a correlation analysis of the collected sense-annotated data. Lastly, we evaluate three baseline WSD models and report on the insights gained. We make both dataset variants publicly available,[1] in the hope that it will facilitate further research in computational semantics for Croatian language.

The rest of the paper is organized as follows. In Section 2, we describe the manual construction of the sense-annotated dataset for Croatian, while in Section 3 we analyze the obtained annotations and we perform a correlation analysis. Section 4 presents the WSD baselines and their evaluation, followed by a discussion. Finally, Section 5 concludes the paper and outlines future work.

2. Building the Cro36WSD Dataset

In the following subsections we explain what data we used (corpora and sense inventory), how we annotated the dataset, and what dataset variants we created.

2.1. Corpus and Sense Inventory

We obtained the data for our dataset from Croatian web corpus hrWaC[2] (Ljubešić and Klubička, 2014), containing 1.9M tokens, annotated with lemma, morphosyntax, and dependency syntax tags.

For the sense inventory, we considered two options: the Croatian WordNet (Raffaelli et al., 2008) and a Croatian machine-readable dictionary (MRD) compiled by Anić (2003).[3] As our previous research has shown that CroWN is fairly incomplete with respect to the sense coverage of polysemous words (Alagić and Šnajder, 2015), we decided to go for the second option. However, larger coverage of MRD does not automatically warrant the appropriateness of the sense inventory. We therefore decided to use the MRD merely as our starting point, subject to further revisions. For the final versions of the sense inventory, we introduced around 80 changes, mostly adding missing senses and modifying other ones, but also discarding unused senses and splitting the overly general ones.

[1] http://takelab.fer.hr/cro36wsd

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

[3] http://hjp.novi-liber.hr/index.php

To select the words for our lexical sample, we have extracted all the polysemous words from the MRD, excluded ones with a frequency lower than 1000, and finally hand-picked 36 words: 12 adjectives, 12 nouns, and 12 verbs. To some extent, we have tried to keep the set of words balanced with respect to both their frequencies and levels of polysemy.
For each of the 36 words, we sampled 300 sentences from hrWaC, which amounts to 10,800 word instances (words and their contexts). Note that for each word in the dataset we meet the criterion for the recommended number of instances per word, namely $75 + 15 \cdot n$, where n denotes the number of word senses (Edmonds and Cotton, 2001).

2.2. Annotation Task

The annotation task was set up rather straightforwardly: each instance to be annotated comprised a sentence containing a target polysemous word, a list of its senses (along with their glosses and usage examples), and an additional "none of the above" (NOTA) option. The annotators were instructed to select all the senses (i.e., multi-label annotation) they deem appropriate for the given word, considering its context. They were also instructed to select the NOTA option in case of an invalid instance, which could occur because of the incorrect lemmatization or spelling errors.
For the semantically opaque contexts (i.e., metaphors and idioms), we have identified three cases:

- The figurative meaning of the word in such context is explicitly listed in the sense inventory. In this case, the annotators were instructed to select it;
- The figurative meaning of the word in such context is not listed in the sense inventory, but the literate one is. Moreover, such meaning fits the predicate frame instantiated in the sentence. In this case, the annotators were instructed to select the literate word meaning;
- Similar to the previous case, except that the literate meaning does not fit the predicate frame. In this case, the annotator wee instructed to select the NOTA option.

Complete annotation was carried out using an online annotation tool (developed in-house) in order to make the annotation both simpler and less time-consuming.

2.3. Annotation Workflow

We divide the dataset annotation into four rounds: (1) a preliminary annotation round, (2) a calibration annotation round, (3) the main annotation round, and (4) an adjudication annotation round. By having four annotation rounds instead of a single one, we are able to revise our sense inventory a couple of times before proceeding with the main annotation round. This results in better and less noisy annotations.

(1) Preliminary annotation round. First, we have randomly sampled 80 instances of each word from the corpus. We then asked four annotators to annotate them and note each of the "problematic" ones along the way. This step served two purposes: we got useful insights into the appropriateness of the sense inventory and we got the opportunity to improve our annotation tool based on annotators' feedback. Taking into account their comments, we revised the initial sense inventory. This round required around 16 person-hours of annotation effort.

(2) Calibration annotation round. Prior to carrying out the main annotation round, we wanted to make sure that our sense inventory is of sufficient coverage and quality and that the annotators have a good understanding of the annotation guidelines. To this end, we asked five new annotators to label a single calibration set. The set contained 25 words that we deemed most problematic, six instances per each. After the annotation was completed, we discussed with the annotators all the systematic annotation errors we have identified. In parallel, we have once more used the obtained insights to revise the sense inventory. Effort for completing this round was around three person-hours.

(3) Main annotation round. As mentioned earlier, we decided to obtain 300 instances for each of the 36 words (10,800 in total) and have them annotated. However, to obtain more robust annotations, we adopted a double-annotation scheme. This increased the number of instance annotations to 21,600, which in turn amounts to 4,320 instances per annotator. Finally, to minimize biases that might be introduced by some annotator pairings, we have distributed the instances uniformly across all possible annotator pairs. We recorded a total effort of 71 person-hours needed to obtain these 21,600 annotations.

(4) Adjudication annotation round. In this last annotation round, annotators were instructed to re-evaluate their sense labels. However, they were asked to do so only for the instances on which they disagreed with the other annotator assigned to the same instance. On each such instance, they were presented with the senses initially selected by the other annotator. This step was primarily meant to rule out systematic mistakes or slips, as it is unlikely that both annotators would make the same mistake in labeling an instance.[4] Adjudication took around 25 person-hours.

2.4. Cro36WSD Variants

To obtain a gold standard sample from data annotated by multiple annotators, their annotations must somehow be aggregated. One common approach is to label each instance with the sense most frequently selected by the annotators (majority voting). Note, however, that this is not possible within a multi-label setup, nor for cases when we have only two annotations per instance. For this reason, we decided to use the following two multi-label aggregation schemes, producing two dataset variants:

- Cro36WSD-M – a *multi-label* variant of the dataset, in which the final label of each instance is obtained as a union of all annotators' labels;
- Cro36WSD-S – a *single-label* variant of the dataset, in which the final label of each instance is obtained as an intersection of all annotators' labels. Instance is retained if the resulting label set is a singleton set, otherwise it is discarded from the dataset.

[4] Even though some disagreements might not be resolved this way, we adopted this strategy due to time and resource constraints.

Combination	Dataset variations	
	Cro36WSD-M	Cro36WSD-S
S_1 & S_2	$S = S_1 \cup S_2$	$S = S_1 \cap S_2$, $\lvert S\rvert = 1$
S_1 & NOTA	$S = S_1$	$S = \emptyset$
NOTA & NOTA	NOTA	NOTA

Table 1: Cro36WSD variants obtained with different label aggregation schemes.

The relationship between these two variants mirrors the one between recall and precision – Cro36WSD-M provides more diverse and lenient annotations, whereas Cro36WSD-S offers a more strict, and thus a more reliable set of annotations (with 774 instances being discarded). A more formal overview of the introduced label aggregation schemes is given in Table 1.

3. Annotation Analysis

In this section, we first present the annotation statistics, including the correlations among the relevant variables. We use the Pearson correlation coefficient r and provide two-tailed p-values for the significance of correlation. Annotation data statistics are given in Table 2 and the correlation figures are given in Table 3.

Average number of NOTA labels. The average number of NOTA labels per word is 25, which seems reasonable taking into account the dataset size. However, there are a few outliers. In the annotations of words *tući* (to beat), *nastaviti* (to continue), and *dom* (home), there are around five times more NOTA labels than the average. The reason behind this mostly lies in unforeseeable systematic lemmatization errors. More precisely, some words were incorrectly lemmatized to a lemma that matches one of the lemmas in our lexical sample, thus introducing an erroneous instance eventually labeled as NOTA. Exception to this rule is the word *nastaviti* (to continue), where NOTA labels possibly occurred due to the overlapping senses, which are difficult to discern.

Sense distribution skewness. We characterize the sense distribution skewness with the sense distribution entropy (E) – the greater the entropy, the less skewed the distribution. We use the standard Shannon's entropy and normalize it with the maximum possible entropy for the given number of senses (i.e., with the entropy of a uniform sense distribution). The word with the lowest entropy, that is, with the most skewed sense distribution is *oprati* (to wash), which was labeled with the same sense 517 out of 561 times.

Average annotation time. As mentioned in Section 2, we noted a total annotation effort of 71 person-hours for the main annotation round. This gives an average per-instance annotation time (AAT) of 12 seconds, which is significantly lower than the one-minute-per-instance estimate reported by Edmonds (2000). Currently we are not aware of the reason behind this considerable difference, but we hypothesize that this could be due to the convenience of our annotation tool.

Inter-Annotator Agreement. We measure the inter-annotator agreement (IAA) using Cohen's κ coefficient (Cohen, 1960). To work around the problem of calculating IAA on multi-label annotation data, we decided to calculate agreement only on single-label instance annotations. Note that the so-obtained IAA is indicative of the overall agreement, as merely 635 out of 21,600 ($\sim 3\%$) instance annotations were multi-label.

We first calculate the agreement for each word separately by averaging the agreements of each annotator pair on their respective portion of that word's instances. We report the micro-averaged IAA for all words in Table 2. By averaging the per-word agreements, we obtain the total IAA score of 0.877, which, according to Landis and Koch (1977), is considered almost a perfect agreement. Considering that the IAA score prior to adjudication was 0.720 (cf. Section 2), the additional adjudication round proved to be a reasonable step to take.

The word with the highest IAA score in our lexical sample is *odlikovati* (to award), with a score of 1. Even though we expected a high agreement on this word, due to its few and very well-distinguishable senses, we still find the perfect agreement rather surprising.

On the other hand, word *pronaći* (to find) was the most difficult one to annotate (IAA of 0.638). We presume that the annotators did not fully grasp the sense definitions, and therefore often confused one sense with another.

Discussion. We observe a significant correlation (r=0.740) between word's AAT and its level of polysemy. It therefore comes as no surprise that highly polysemous words, such as *star* (old), *pojas* (belt), and *pasti* (to fall), are associated with the longest average annotation time. The same observation has been made by Kapelner et al. (2012), but within a crowdsourcing annotation context. Note, however, that AAT does not depend solely on the number of word senses. Some less polysemous words also take a lot of time to be annotated – the word *normalan* (normal) proved to be quite problematic for the annotators, while having as few as three senses. We hypothesize that this might be due to the senses being too fine-grained, and therefore hardly discernible in everyday use.

Another significant correlation (r=0.475) is the one between the AAT and the skewness of the sense distribution. While annotating the words with fairly skewed sense distributions (e.g., *oduzeti* (to take away)), annotators would often encounter a lengthy sequence of instances in which the word always bears the same sense. Additionally, due to already discussed skewness of word sense distributions, only a handful of most frequent senses occurred in such sequences. For this reason, annotators did not need to pay as much attention (nor spend a lot of time) considering other senses, as they would otherwise.

Interestingly enough, we do not observe a significant correlation between word's AAT and IAA (r= –0.271), even though we find it quite reasonable to presume that shorter AAT implies easier annotation, and in turn higher IAA.

4. Baseline WSD Experiments

Baseline WSD models. We use three baseline models for our experiments. The first one is a standard Most Frequent Sense (MFS) classifier, which has been proven to be a competitive WSD baseline (Agirre and Edmonds, 2007). To handle the multi-label classification, we adapt MFS in such

Word (hr)	Word (en)	POS	Freq.	S	Sense distribution	NOTA	E	IAA κ	AAT
aktivan	active	A	85281	5	173 / 283 / 98 / 9 / 39	20	0.754	0.806	14.0
lak	easy	A	15424	7	93 / 1 / 56 / 12 / 308 / 78 / 35	31	0.728	0.859	14.9
mrtav	dead	A	14252	4	474 / 50 / 52 / 20	13	0.500	0.865	8.2
normalan	normal	A	109594	3	410 / 114 / 93	5	0.655	0.714	15.2
oštar	sharp	A	27221	8	29 / 37 / 398 / 27 / 41 / 48 / 11 / 21	18	0.634	0.902	18.4
pažljiv	careful	A	5950	2	497 / 92	26	0.537	0.677	8.2
pokvaren	broken	A	5789	5	169 / 35 / 298 / 39 / 77	1	0.733	0.995	10.2
prljav	dirty	A	14245	3	261 / 317 / 10	22	0.642	0.818	8.1
siguran	sure	A	222067	4	154 / 161 / 264 / 23	11	0.781	0.923	12.6
star	old	A	350446	8	121 / 64 / 323 / 41 / 63 / 11 / 19 / 3	22	0.722	0.906	20.3
vanjski	outer	A	196993	3	210 / 61 / 343	0	0.665	0.981	7.2
visok	high	A	439729	5	100 / 15 / 92 / 288 / 106	33	0.821	0.839	17.8
dom	home	N	268586	7	239 / 22 / 40 / 198 / 45 / 1 / 0	91	0.720	0.762	12.7
godina	year	N	4530788	5	329 / 204 / 51 / 20 / 0	1	0.575	0.906	11.4
okvir	frame	N	141862	4	73 / 505 / 3 / 19	2	0.347	0.963	7.3
pojas	belt	N	33805	10	65 / 20 / 11 / 120 / 196 / 53 / 2 / 10 / 122 / 6	57	0.801	0.765	21.2
pokrivač	cover	N	8871	3	261 / 313 / 9	24	0.645	0.900	7.0
povreda	injury	N	18179	2	143 / 457	5	0.539	0.963	10.1
publika	audience	N	191548	2	359 / 288	1	0.635	0.822	11.1
rezerva	reserve	N	12921	4	387 / 146 / 5 / 36	29	0.610	0.948	8.8
težina	weight	N	68112	5	13 / 133 / 7 / 415 / 43	0	0.510	0.961	9.3
trag	trace	N	66790	4	70 / 173 / 289 / 57	38	0.836	0.820	13.0
vatra	fire	N	45943	8	283 / 108 / 2 / 2 / 3 / 23 / 3 / 186	9	0.590	0.968	9.7
zvanje	vocation	N	12992	4	442 / 143 / 3 / 30	6	0.496	0.957	8.3
brusiti	to rasp	V	1514	5	47 / 216 / 24 / 33 / 306	8	0.694	0.941	16.9
gorjeti	to burn	V	3126	4	507 / 67 / 13 / 24	29	0.474	0.884	9.6
kucati	to knock	V	5368	4	303 / 11 / 242 / 1	45	0.615	0.884	9.0
nastaviti	to continue	V	280913	3	480 / 0 / 26	95	0.438	0.940	7.7
odlikovati	to award	V	15504	2	150 / 450	0	0.512	1.000	4.7
oduzeti	to take away	V	35411	3	557 / 51 / 6	2	0.260	0.896	10.1
oprati	to wash	V	10034	4	517 / 8 / 26 / 10	46	0.368	0.733	8.3
osuditi	to judge	V	32108	3	134 / 396 / 44	26	0.676	0.829	8.1
pasti	to fall	V	168967	13	63 / 83 / 22 / 2 / 8 / 1 / 3 / 29 / 254 / 44 / 3 / 71 / 9	33	0.724	0.882	24.7
poslužiti	to serve	V	54033	3	91 / 425 / 79	6	0.609	0.942	11.6
pronaći	to find	V	225148	3	51 / 548 / 18	3	0.320	0.638	16.4
tući	to beat	V	18729	8	341 / 0 / 4 / 32 / 72 / 11 / 4 / 8	140	0.575	0.989	13.3

Table 2: Annotation data statistics. (S stands for number of senses and E stands for entropy.)

	Freq.	S	E	IAA
S	0.038			
E	– 0.004	0.391		
IAA	0.032	0.064	– 0.012	
AAT	0.040	0.740*	0.475*	– 0.271

Table 3: Pearson correlation coefficient r between various statistics (* marks the significance with $p < 0.01$).

a way that it always predicts the most frequent *set* of senses (labels) in the training set. For the other two baselines we use Support Vector Machine (SVM) classifiers based on different context representations. To handle the multi-label classification, we independently train one binary classifier for each label, obtaining the final set of labels as a union of all labels that were predicted as positive by the respective classifiers (Tsoumakas and Katakis, 2007). We use the freely available LIBSVM library (Chang and Lin, 2011).

Both SVM baselines we use leverage the most simple, word-based context representations: one uses a standard bag-of-words (BoW) context representation, whereas the other one uses skip-gram (SG) vectors (Mikolov et al., 2013). For both representations, the composed context vector is obtained by summing the vectors of the words found in the context. We build the skip-gram vectors from hrWaC using the `word2vec`[5] tool. We use 250 dimensions, negative sampling parameter set to 5, minimum frequency set to 100, and no hierarchical softmax.

We randomly split the dataset into a training and a test set: for each of the selected words, we use around 200 instances for training (the number varies across dataset variants and words) and 100 instances for testing. We optimize the hyperparameters of the SVM using a 5-fold cross-validation (optimizing for micro-F1 score) on the training set. By treating each label in a multi-label output as a separate prediction, our micro-F1 score is the computed the same for both the single- and multi-label classification. We report micro-F1 scores of the baseline models in Table 4.

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

			Cro36WSD-M			Cro36WSD-S		
Word (hr)	Word (en)	POS	MFS	BoW	SG	MFS	BoW	SG
aktivan	active	A	0.500	0.539	**0.606**	0.430	0.530	**0.610**
lak	easy	A	0.530	**0.845**	0.802	0.510	0.670	**0.710**
mrtav	dead	A	0.800	**0.804**	0.787	0.830	**0.850**	**0.850**
normalan	normal	A	0.750	0.750	**0.769**	0.740	0.740	**0.780**
oštar	sharp	A	0.670	0.739	**0.849**	0.640	0.690	**0.850**
pažljiv	careful	A	**0.900**	0.890	0.889	0.860	0.860	**0.910**
pokvaren	broken	A	0.470	0.750	**0.785**	0.510	0.660	**0.790**
prljav	dirty	A	0.520	0.667	**0.779**	0.510	0.780	**0.860**
siguran	sure	A	0.490	0.567	**0.571**	0.400	**0.500**	0.470
star	old	A	0.570	0.654	**0.687**	0.510	**0.550**	0.520
vanjski	outer	A	0.520	0.880	**0.892**	0.600	0.900	**0.940**
visok	high	A	0.470	0.600	**0.654**	0.500	0.580	**0.640**
Average:			0.599	0.724	**0.756**	0.587	0.692	**0.744**
dom	home	N	0.450	0.644	**0.747**	0.250	0.430	**0.570**
godina	year	N	0.650	**0.658**	0.594	0.490	**0.560**	0.490
okvir	frame	N	0.830	**0.850**	0.842	0.870	0.880	**0.900**
pojas	belt	N	0.420	**0.857**	0.842	0.270	0.560	**0.690**
pokrivač	cover	N	0.550	0.787	**0.824**	0.530	0.840	**0.930**
povreda	injury	N	0.760	0.810	**0.864**	0.810	0.860	**0.920**
publika	audience	N	0.690	0.684	**0.785**	0.510	0.620	**0.700**
rezerva	reserve	N	0.680	0.802	**0.839**	0.570	0.700	**0.820**
težina	weight	N	0.700	0.711	**0.845**	0.670	0.730	**0.760**
trag	trace	N	0.590	0.726	**0.783**	0.410	0.630	**0.650**
vatra	fire	N	0.520	**0.915**	0.909	0.430	0.840	**0.910**
zvanje	vocation	N	0.790	0.880	**0.913**	0.720	0.850	**0.890**
Average:			0.636	0.777	**0.816**	0.544	0.708	**0.769**
brusiti	to rasp	V	0.480	**0.815**	0.807	0.540	0.790	**0.810**
gorjeti	to burn	V	0.840	0.842	**0.867**	0.820	0.820	**0.830**
kucati	to knock	V	0.530	0.885	**0.888**	0.510	**0.950**	0.920
nastaviti	to continue	V	0.810	0.901	**0.936**	0.830	0.920	**0.930**
odlikovati	to award	V	0.730	0.900	**0.970**	0.730	0.900	**0.970**
oduzeti	to take away	V	0.930	0.930	**0.949**	**0.940**	**0.940**	**0.940**
oprati	to wash	V	0.890	0.890	**0.899**	**0.900**	**0.900**	0.770
osuditi	to judge	V	0.660	0.888	**0.920**	0.650	**0.870**	**0.870**
pasti	to fall	V	0.380	**0.933**	0.769	0.470	0.520	**0.550**
poslužiti	to serve	V	0.690	0.750	**0.806**	0.670	0.730	**0.770**
pronaći	to find	V	**0.930**	**0.930**	**0.930**	**0.890**	**0.890**	**0.890**
tući	to beat	V	0.610	0.883	**0.884**	0.550	0.860	**0.890**
Average:			0.707	0.879	**0.885**	0.708	0.841	**0.845**
Total average:			0.647	0.793	**0.819**	0.613	0.747	**0.786**

Table 4: Baseline WSD models micro-F1 scores on Cro36WSD-M and Cro36WSD-S.

Discussion. By analyzing the average WSD performance, we notice that both BoW and SG models outperform the MFS model by a significant margin, on both Cro36WSD-M and Cro36WSD-S dataset variants. In addition, SG model proves to be numerically better than the BoW model. Note, however, that MFS classifier still performs competitively (as noted by Agirre and Edmonds (2007)), which is mostly due to the inherent skewness of word sense distributions. Consequently, MFS performs best on words with very skewed sense distributions, namely *oduzeti* (to take away), *pronaći* (to find), and the like. This correlation is significant for both Cro36WSD-M (r=–0.604) and Cro36WSD-S (r=–0.583).

We also analyze the average WSD performance across POS tags and models. In all but one case, models exhibit the best performance on verbs, followed by nouns, and adjectives. While we can exclude the hypothesis that these differences stem from the differences in frequency or polysemy level (as our lexical sample is balanced across these variables), we leave a more detailed investigation for future work. Interestingly enough, the only case where this observation does not hold is on Cro36WSD-S, where MFS model works better on adjectives (0.587) than on nouns (0.544).

To test the statistical significance of our results, for each of the two datasets we test the difference between micro-F1 scores for the different models across all words in the lexical sample. To this end, we use the two-tailed matched pairs t-test. For both dataset variants, the differences between any pair of models are statistically significant ($p < 0.01$).

Taking a look at the best-performing model, namely SG, we find the word with the highest F1-score to be *odlikovati* (to award), in both single- and multi-label dataset variants. This is because of the small number of well-distinguishable senses in the sense inventory, which is in line with high IAA for this word (cf. Section 2). However, the word with the lowest F1-score (again for both dataset variants) is *siguran* (sure), despite of its very high IAA (0.932). We hypothesize that this is due to its contexts being less informative and not discriminative enough to disambiguate its meaning. We leave a more detailed investigation for future work.

5. Conclusion

Cro36WSD is a medium-sized lexical sample for Croatian word sense disambiguation (WSD). It comprises 12 adjectives, 12 nouns, and 12 verbs, balanced across both frequency bands and polysemy levels. We adopted a multi-label annotation scheme, which allowed us to generate two different dataset variants using different label aggregation schemes. To obtain high-quality annotations, we annotated the sample in four rounds, and achieved an overall inter-annotator agreement of 0.877. We observe significant correlations between the average annotation time and both the level of polysemy and the sense distribution skewness. We report on the performance of three baseline WSD systems on this dataset. We make both datasets publicly available.

6. Acknowledgements

This work has been fully supported by the Croatian Science Foundation under the project UIP-2014-09-7312.

7. Bibliographical References

Agirre, E. and Edmonds, P. (2007). *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media.

Agirre, E., De Lacalle, O. L., Fellbaum, C., Marchetti, A., Toral, A., and Vossen, P. (2009). SemEval-2010 task 17: All-words word sense disambiguation on a specific domain. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 123–128.

Alagić, D. and Šnajder, J. (2015). Experiments on active learning for Croatian word sense disambiguation. In *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing, BSNLP 2015*, pages 49–58, Hissar, Bulgaria.

Anić, V. (2003). *Veliki rječnik hrvatskoga jezika*. Novi Liber.

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

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Edmonds, P. and Cotton, S. (2001). Senseval-2: overview. In *Proceedings of SensEval-2*, pages 1–5, Toulouse, France.

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

Edmonds, P. (2000). Designing a task for Senseval-2.

Erk, K. and McCarthy, D. (2009). Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 1 of *EMNLP '09*, pages 440–449, Singapore.

Kapelner, A., Kaliannan, K., Schwartz, H. A., Ungar, L. H., and Foster, D. P. (2012). New insights from coarse word sense disambiguation in the crowd. In *COLING (Posters)*, pages 539–548.

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

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

Manandhar, S., Klapaftis, I. P., Dligach, D., and Pradhan, S. S. (2010). SemEval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 63–68.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119, Nevada, USA.

Moro, A. and Navigli, R. (2015). Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. *Proceedings of SemEval-2015*.

Navigli, R., Litkowski, K. C., and Hargraves, O. (2007). SemEval-2007 task 07: Coarse-grained english all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 30–35.

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

Raffaelli, I., Tadić, M., Bekavac, B., and Agić, Ž. (2008). Building Croatian wordnet. In *Proceedings of GWC*, pages 349–360, Szeged, Hungary.

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

Stokoe, C., Oakes, M. P., and Tait, J. (2003). Word sense disambiguation in information retrieval revisited. In *Proceedings of ACM SIGIR*, pages 159–166, Toronto, Canada.

Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. *The International Journal of Data Warehousing and Mining*, 2007:1–13.