Graph-Based Induction of Word Senses in Croatian

Marko Bekavac and Jan Šnajder
University of Zagreb, Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
{jan.snajder, marko.bekavac}@fer.hr

Abstract

Word sense induction (WSI) seeks to induce senses of words from unannotated corpora. In this paper, we address the WSI task for the Croatian language. We adopt the word clustering approach based on co-occurrence graphs, in which senses are taken to correspond to strongly inter-connected components of co-occurring words. We experiment with a number of graph construction techniques and clustering algorithms, and evaluate the sense inventories both as a clustering problem and extrinsically on a word sense disambiguation (WSD) task. In the cluster-based evaluation, Chinese Whispers algorithm outperformed Markov Clustering, yielding a normalized mutual information score of 64.3. In contrast, in WSD evaluation Markov Clustering performed better, yielding an accuracy of about 75%. We are making available two induced sense inventories of 10,000 most frequent Croatian words: one coarse-grained and one fine-grained inventory, both obtained using the Markov Clustering algorithm.

Keywords: word sense induction, co-occurrence graph, graph-based clustering, Croatian language

1. Introduction

Word sense disambiguation (WSD) – the task of automatically determining the sense of a polysemous word in context – is considered one of the fundamental problems of natural language processing (Navigli, 2009). WSD can be approached as either *sense labeling* or *sense discrimination*. In the former, words occurrences are assigned sense labels from a predefined sense inventory. In contrast, sense discrimination (Schütze, 1998) addresses a simpler task of differentiating among the different uses of a word, without a reference to a sense inventory.

Related to WSD is the task of word sense induction (WSI). WSI seeks to induce senses of words from unannotated corpora. There are two main approaches to WSI: *context clustering* and *word clustering*. Context clustering exploits the distributional hypothesis (Harris, 1954) to group together the similar usages of a word. In contrast, word clustering groups together different but semantically similar words that pertain to a specific sense (e.g., *money*, *loan*, and *finance* for the financial sense of *bank*). In particular, word clustering based on *co-occurrence graphs* posits that senses of a polysemous word correspond to strongly interconnected components of its co-occurring words.

In this paper, we address the WSI task for Croatian, a resource-scarce South Slavic language. Our aim is to build general-domain sense-induced inventories (in the form of word clusters) that can be readily used for WSD and related tasks. We adopt a word clustering approach using co-occurrence graphs and experiment with a number of graph construction techniques and clustering algorithms. The contribution of our work is twofold: (1) we describe and make publicly available induced sense inventories for Croatian and (2) we evaluate the sense inventories both intrinsically (as a clustering problem) and extrinsically on a WSD task. To the best of our knowledge, this is the first work on WSI for Croatian, and also the first that makes such resources freely available.

The rest of the paper is structured as follows. In the next section, we give a brief overview of the related work. In Section 3 we describe the co-occurrence graphs and the graph-based clustering. In Section 4 we describe the manual annotation of a gold standard dataset for the intrinsic WSI evaluation. Section 5 presents the evaluation results. Section 5 concludes the paper.

2. Related Work

Studies on graph co-occurrence based WSI date back to the work of Widdows and Dorow (2002), whose algorithm clustered the nodes (nouns only) using the similarities between their neighbors. The HyperLex (Véronis, 2004) algorithm exploits the small world property of co-occurrence graphs and extracts the hubs corresponding to word senses. Agirre et al. (2006) optimize the parameters of the HyperLex algorithm and compare it to a modified version of PageRank. Di Marco and Navigli (2013) study WSI in the context of web search result clustering.

The evaluation of WSD and WSI systems has been given a great deal of attention in the literature. Most evaluation methods and resources stem from the SemEval (Senseval) workshops (Agirre and Soroa, 2007; Manandhar and Klapaftis, 2009). WSI systems are typically evaluated on the task of unsupervised WSD, either sense labeling or sense discrimination. In both cases, a sense inventory and a corresponding sense-annotated corpus are required. Unlike previous work, in this paper we focus on an intrinsic evaluation of our WSI system, which does not require a sense-annotated corpus. Additionally, we evaluate our system on the task of unsupervised WSD.

Our work focuses on WSI for Croatian, which has not yet been addressed in the literature. The existing work for Croatian focuses exclusively on the WSD task. Bakarić et al. (2007) analyze the discriminative strength of WSD predictors. Alagić and Šnajder (2015) study the efficacy of active learning for Croatian WSD on a manually annotated lexical sample comprising six polysemous words. Re-

cently, Alagić and Šnajder (2016) presented a larger WSD dataset for Croatian comprising 36 polysemous words.

3. Inducing Co-occurrence Clusters

The induction of words sense clusters comprises two steps: the construction of weighted co-occurrence graphs from the corpus and the graph-based clustering. We next describe these steps in detail.

3.1. Co-occurrence Graphs

The idea behind the co-occurrence graph is to represent a network of words in which semantically related words will have many neighbors in common. If a group of nodes is strongly inter-connected, we may assume that they often appear in the same context. On the basis of the distributional hypothesis, in WSI we assume that different senses of a word are determined by the different contexts. Therefore, if we can separate the neighbors of a polysemous word into strongly-connected clusters, each of them will represent a distinct sense.

Corpus and preprocessing. We extract the co-occurrence graphs from the Croatian web-corpus hrWaC (Ljubešić and Erjavec, 2011), totaling 1.2 billion tokens. The corpus was lemmatized and POS-tagged using the tools from (Agić et al., 2013).

Node filtering. The shape of a co-occurrence graph crucially depends on the size of the corpus: if the corpus is too small, many relations will not be present, and rare senses will not be captured. On the other hand, a large corpus may yield a noisy co-occurrence graph. To account for this, prior to building the co-occurrence graph, we filtered out words that occurred less than 100 times in the corpus. From these, we retained only the nouns, verbs, and adjectives, leaving 160,052 words. We consider a pair of word to be co-occurring if they occur in the same sentence. After we built the graph, we removed the edges between words that co-occurred only once.

Graph weighting. Co-occurrence frequency alone does not provide reliable estimates of word similarities, as a frequent word tends to co-occur with many other words, simply as a result of its high frequency. A proper word association measure should take into account the frequencies of both words as well as the number of their co-occurrences. A number of word association measures has been proposed in the literature (Terra and Clarke, 2003; Pecina, 2010). We consider five well-known association measures: log-likelihood, Dice coefficient, χ^2, z-score, pointwise mutual information, and local mutual information (Evert, 2005). In a preliminary experiment, we evaluated the measures on a dataset of 450 word pairs with human-judged similarity ratings from Janković et al. (2011). In terms of correlation with human-assigned ratings, the Dice coefficient outperformed the other three considered measures (Pearson's correlation coefficient of $r = 0.489$) by a significant margin (the second-highest ranking measure, z-score, reached $r = 0.421$). We therefore chose Dice for weighting the co-occurrence graph.

Edge filtering. After we have obtained a weighted co-occurrence graph, we removed from it the edges whose weight is below an experimentally obtained threshold. To determine the threshold, we started with a low threshold value, and then sampled the edges whose weight is close to the threshold value. If we were not able to identify a semantic relatedness between any of the incident nodes, we would increase the threshold value. This resulted in a final value of 10^{-4}, retaining 79.8% of the graph edges. The final graph has 106,476 nodes and 19,846,760 weighted edges.

3.2. Clustering

Subgraph extraction. Following previous work on co-occurrence based WSI (Dorow and Widdows, 2003), we chose not to induce the senses on the complete graph. Instead, we induce the senses for each target word separately, by clustering the subgraph centered around the target word. Besides being computationally more efficient, clustering on a per-word basis allows for overlapping word clusters.[1] The subgraph consists of all first- and second-degree neighbors of the target word. However, to keep the size of the subgraph manageable, we increment the edge weight threshold as we move away from the target word. The threshold value is not fixed in advance, rather it is adjusted to yield a subgraph whose size is within the minimum and maximum number of nodes. We set the minimum number of nodes to 20 and the maximum number of nodes to 40.

The average number of nodes in a subgraph is 25.3, while the average number of edges is 241. Note that this graph is less sparse than the complete graph we started from.

Graph-based clustering. After obtaining the subgraph for each target word, we use graph-based clustering algorithms to cluster the co-occurring words. We experimented with a number of clustering algorithms: B-MST (Di Marco and Navigli, 2011; Di Marco and Navigli, 2013), SquaT++ (Di Marco and Navigli, 2013), Chinese Whispers (Biemann, 2006), HyperLex (Véronis, 2004), PageRank (Agirre et al., 2006), HITS (Gibson et al., 1998), and Markov Clustering (van Dongen, 2000). Our experiments revealed that Markov Clustering and Chinese Whispers outperform the other considered algorithms. In what follows, we restrict our attention to these two algorithms.

The Chinese Whispers (CW) algorithm works in a bottom-up fashion: starting from a configuration in each node constitutes a singleton cluster, the algorithm iteratively builds larger clusters by randomly choosing nodes and assigning them to a cluster to which they have the strongest average connection. The procedure is repeated until it converges or reaches a predefined number of iterations. Although there is no convergence guarantee, clustering usually stabilizes after a handful of iterations. Thus, the CW algorithm is essentially parameter-free and there is no need for parameter tuning.

Markov Clustering (MCL) models stochastic flows in graphs as random walks over the nodes, with the probability of traversing a particular edge being reciprocal to the

[1] The alternative would be to cluster only the words that occur in the same paragraph as the target word, as proposed by Véronis (2004).

avion (*airplane*)	otkazati (*cancel*)
cilj (*goal*)	politika (*politics*)
dokument (*document*)	pomagalo (*tool*)
glazbalo (*musical instrument*)	poslovanje (*business conduct*)
isprava (*decree*)	predmet (*object*)
kamata (*interest rate*)	rad (*work*)
kontrolan (*control*)	snimanje (*recording*)
laboratorij (*laboratory*)	sredstvo (*means*)
mjerenje (*measuring*)	sustav (*system*)
ordinacija (*doctor's office*)	tonski (*tonal*)
zvuk (*sound*)	

Table 1: Gloss words for the word "instrument" (*instrument*).

edge's weight. A random walk is likely to stay within a strongly-connected component, which can then be considered a cluster. MCL uses two operators: expansion and inflation. Expansion coincides with matrix squaring while inflation consists of taking the Hadamard power of a matrix followed by a diagonal scaling step. We used only one parameter, inflation value, which does not explicitly determine the number of clusters, but rather affects the cluster granularity (higher inflation value yields finer granulation). As suggested by van Dongen (2000), we set the other parameters to their defaults.

4. Gold Standard Sense Annotation

Previous work uses sense-annotated corpora for WSD-based evaluation of WSI systems. We argue that there are three main shortcomings of such an evaluation. First, compiling a sense-annotated sample is tedious and labor-intensive. Secondly, the evaluation is tied to a sense inventory, and hence inherits all the problems associated with fixed sense inventories. Finally, WSD-based evaluation does not at all differentiate between the quality of the induced sense inventory and the efficiency of the WSD procedure.

To account for the above deficiencies, we opt for a different route. Our Gold Standard compilation works as follows. First, for each of the ambiguous word, we constructed a set of words from its gloss. We then asked the annotators to cluster these words into sense clusters, giving them as much freedom as possible, and allowing for one word to appear in multiple clusters (soft clustering). Based on the similarity between the so-obtained clusterings, we selected a subset of annotators with the highest agreement. We then aggregated their clusterings into a single soft clustering. Finally, we transformed the so-obtained soft clustering into a hard clustering. We next describe these steps in more detail.

Word clusters. We sampled 45 polysemous words (15 nouns, adjectives, and verbs each) from a Croatian machine-readable dictionary (Anić, 2003). To compensate for the effects of word frequency, within each part-of-speech, we chose 5 words from each the upper (ranks above 100), middle (ranks around 1000), and lower (ranks below 10,000) frequency band in hrWaC. For each word, we extracted from the glosses of all its senses all nouns, adjectives, and verbs. This gives us a set of gloss words for each of the 45 words. We omitted the words that appear in more than one gloss to make it easier to perform WSD evaluation (cf. Section 5). As an example, Table 1 shows the gloss words for the ambiguous word "instrument" (instrument), across all its senses from the machine readable dictionary.

avion	cilj	glazbalo	isprava
kontrolan	kamata	pomagalo	kontrolan
laboratorij	otkazati	predmet	pomagalo
mjerenje	politika	rad	poslovanje
ordinacija	poslovanje	snimanje	predmet
otkazati	rad	sredstvo	rad
pomagalo	sredstvo	sustav	sredstvo
predmet	sustav	tonski	
rad		zvuk	
snimanje			
sredstvo			
sustav			

Table 2: Annotator-clustered gloss words for senses of "instrument".

Annotation. We next asked 10 annotators to cluster the sets of gloss words for each of the 45 words. The annotators were instructed to group the words so that they represent the distinct senses of the words. A gloss word could be used in more than one cluster or could be ignored. The annotators worked independently of each other. Table 2 shows one of the clusterings obtained for the word "instrument". Note that the clusters share some of the gloss words.

Annotator agreement. Having obtained 10 annotations for each word, we next transformed the clusterings for each word into a stochastic matrix, encoding the probabilities of two words appearing in the same cluster. We calculated the disagreement between annotators as the Jensen-Shannon divergence.

For each word, we chose clusterings of six annotators with the highest agreement and computed the gold standard clusters for that word as the average of the six human-annotated clusterings. We used the Hierarchical Agglomerative Clustering (HAC) algorithm for each target word to find the set of annotators with the highest agreement. The result of this procedure is a set of 45 soft clusterings.

Hard clustering. Finally, we ran the Hierarchical Agglomerative Clustering (HAC) algorithm with average linkage on the probability (similarity) matrix of each word to obtain the hard clusters. The number of clusters was set to the integer closest to average number of senses identified by the annotators for that particular target word. The net result are word clusters for each of the 45 polysemous words, where each clustering represents the senses of that particular word as perceived by human annotators. The average number of clusters is 3.5.

Table 3 shows the final gold clusters for the example word "instrument", while Table 4 shows the clusters that correspond to the senses listed in the machine-readable dictionary. The annotators found a fewer number of senses, which to a certain extent can be derived by merging some of the lexicon senses.

cilj	dokument	glazbalo	avion
kamata	isprava	snimanje	
otkazati	kontrolan	tonski	
politika	laboratorij	zvuk	
poslovanje	mjerenje		
sredstvo	ordinacija		
sustav	pomagalo		
	predmet		
	rad		

Table 3: Gold standard clustering for "instrument".

laboratorij	avion	glazbalo	dokument	kamata	cilj
ordinacija	kontrolan	predmet	isprava	poslovanje	politika
pomagalo	mjerenje	tonski		sredstvo	
rad	otkazati	zvuk			
	snimanje				
	sustav				

Table 4: Clusters (glosses) for each of the dictionary senses.

5. Evaluation

Cluster evaluation. We evaluate the clustering by comparing the induced clusters against gold clusters using the normalized mutual information (NMI), purity, and the F1-measure. Note that the words from gold clusters will not appear in the induced clusters if they did not exist in the co-occurrence subgraph. To account for this, we assign each such word to the induced cluster to which it is most similar to in terms of single linkage computed on the full co-occurrence graph.

We consider three clustering variants: (1) CW clustering, (2) MCL clustering with the inflation parameter set to optimize the NMI score, and (3) MCL clustering with the inflation parameter set to the value for which the average number of clusters is equal to the average number of gold standard clusters. Variants (1) and (2) yield clusters of average size 6.6 and 6.0, respectively. Thus, the first two variants may be considered fine-grained clusters, while the third one is more coarse-grained.

The MCL inflation parameters are 16.5 for the optimum point and 1.8 for a more coarse-grained clustering that produces an average of 3.5 clusters.

Results are shown in Table 5. Chinese Whispers outperform both variants of MCL in terms of NMI score, while it underperforms in terms of purity, where fine-grained MCL is the best. All of the clustering variants scored roughly the same in terms of the F1-measure. Overall, the scores indicate that the induced clusters only loosely match the human-annotated sense clusters.

WSD evaluation. In addition to the cluster-based evaluation described above, we evaluate WSI in terms of its performance on an unsupervised WSD task.

Our focus is not on the performance of the WSD itself; we rather want to illustrate how the information gained from WSI can be used and whether it can match human-compiled sense inventory in the disambiguation task.

For the WSD-based evaluation of the induced sense inventories, we chose nine words from the 45 in our sample, three for each part-of-speech. For each of the nine words, we sampled five sentence from hrWaC, and annotated their senses using the sense inventory from (Anić, 2003). To disambiguate a word in context, we used a variant of the Lesk algorithm (Lesk, 1986). First, we map each of the context words to the WSI cluster to which it is most similar (in terms of the Dice coefficient on the full co-occurrence graph). We experimented with different methods of connecting the context words to the clusters, and single linkage proved to work best: the context word got mapped to the WSI cluster to which it had a single highest-weight connection. Next, we chose the cluster that has received the most hits relative to its size (we experimented with a couple of other but less successful schemes). Finally, we choose the sense for which the relative size of the overlap between gloss and WSI cluster is maximized. Note that, unlike in the original Lesk algorithm, we match the words to senses via co-occurrence clusters. This maximizes the chance of a correct match because clusters are generally larger and more similar to context words than glosses.

Algorithm	Avg. number of clusters	NMI	Purity	F1
Chinese Whispers	6.6	**64.3**	4.4	40.3
MCL (fine)	6.0	28.4	**60.2**	38.8
MCL (coarse)	3.5	18.0	55.3	**42.8**

Table 5: Cluster evaluation on gold word groups.

Algorithm	Accuracy
Chinese Whispers	17.6
MCL (fine)	63.0
MCL (coarse)	**75.7**
Lesk	49.5
MFS	68.8

Table 6: WSD accuracy of the induced sense inventories.

The results are shown in Table 6. We consider two baselines: the Lesk algorithm and the most frequent sense (MFS) baseline. Coarse-grained MCL outperformed the baselines by a large margin. It is also the only method that outperformed the very competitive MFS baseline. Lesk algorithm performed worse than MFS, indicating that WSI clusters do a good job in bridging the lexical gap between contexts and glosses. CW clusters performed poorly, despite the good results in terms of cluster evaluation.

We hypothesize that this discrepancy can be traced back to the differences between the WSI clustering task and the WSD task: in the WSD task, being able to model the most frequent sense is rather important, while in the WSI clustering task all senses are given equal importance. It seems that the Chinese Whispers has succeeded in finding many of the less represented senses (which the annotators also found), but failed to distinguish between the dominant and the non-dominant ones. In contrast, both variants of the MCL algorithm succeeded in modeling the dominant senses. This is also evident from the obtained cluster sizes: the Chinese Whispers had more evenly sized clusters, while MCL almost always had one more prominent cluster.

6. Conclusion

We presented our experiments on co-occurrence graph word sense induction (WSI) for Croatian. We described an intrinsic evaluation setup, in which Chinese Whispers algorithm outperformed Markov Clustering. In contrast, in word sense disambiguation (WSD) evaluation, Markov Clustering emerged as the winner, with a rather good accuracy of about 75%. Investigating the relation between intrinsic and extrinsic WSI performance is part of future work.

We are making the induced sense inventories for 10,000 most frequent Croatian words freely available.[2] The dataset contains clusterings obtained by using the MCL algorithm with two inflation parameters, one yielding a coarse-grained sense clusters, and the other yeilding find-grained sense clusters.

7. Acknowledgments

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

8. Bibliographical References

Agić, Ž., Ljubešić, N., and Merkler, D. (2013). Lemmatization and morphosyntactic tagging of Croatian and Serbian. In *Proceedings of ACL*.

Agirre, E. and Soroa, A. (2007). SemEval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 7–12. ACL.

Agirre, E., Martínez, D., de Lacalle, O. L., and Soroa, A. (2006). Two graph-based algorithms for state-of-the-art wsd. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 585–593. ACL.

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. ACL.

Alagić, D. and Šnajder, J. (2016). Cro36WSD: A lexical sample for croatian word sense disambiguation. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*, page In press.

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

Bakarić, N., Njavro, J., and Ljubešić, N. (2007). What makes sense? Searching for strong wsd predictors in Croatian. In *Digital Information and Heritage*. Odsjek za informacijske znanosti Filozofskog fakulteta u Zagrebu.

Biemann, C. (2006). Chinese Whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the first workshop on graph based methods for natural language processing*, pages 73–80. ACL.

Di Marco, A. and Navigli, R. (2011). Clustering web search results with maximum spanning trees. In *AI* IA 2011: Artificial Intelligence Around Man and Beyond*, pages 201–212. Springer.

Di Marco, A. and Navigli, R. (2013). Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, 39(3):709–754.

Dorow, B. and Widdows, D. (2003). Discovering corpus-specific word senses. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2*, pages 79–82. ACL.

Evert, S. (2005). *The statistics of word cooccurrences: word pairs and collocations*. Ph.D. thesis, IMS, Universität Stuttgart.

Gibson, D., Kleinberg, J., and Raghavan, P. (1998). Inferring web communities from link topology. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space—structure in hypermedia systems: links, objects, time and space—structure in hypermedia systems*, pages 225–234. ACM.

Harris, Z. S. (1954). Distributional structure. *Word*.

Janković, V., Šnajder, J., and Bašić, B. D. (2011). Random indexing distributional semantic models for Croatian language. In *Text, Speech and Dialogue*, pages 411–418. Springer.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM.

Ljubešić, N. and Erjavec, T. (2011). hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *Text, Speech and Dialogue*, pages 395–402. Springer.

Manandhar, S. and Klapaftis, I. P. (2009). Semeval-2010 task 14: evaluation setting for word sense induction & disambiguation systems. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 117–122. ACL.

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

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language resources and evaluation*, 44(1-2):137–158.

Schütze, H. (1998). Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123.

Terra, E. and Clarke, C. L. (2003). Frequency estimates for statistical word similarity measures. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 165–172. ACL.

van Dongen, S. M. (2000). Graph clustering by flow simulation.

Véronis, J. (2004). Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252.

Widdows, D. and Dorow, B. (2002). A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. ACL.

[2] `http://takelab.fer.hr/crowsi`