Leveraging Lexical Substitutes for Unsupervised Word Sense Induction

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

Sebastian Padó
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
pado@ims.uni-stuttgart.de

Abstract

Word sense induction is the most prominent unsupervised approach to lexical disambiguation. It clusters word instances, typically represented by their bag-of-words contexts. Therefore, uninformative and ambiguous contexts present a major challenge. In this paper, we investigate the use of an alternative instance representation based on *lexical substitutes*, i.e., contextually suitable, meaning-preserving replacements. Using lexical substitutes predicted by a state-of-the-art automatic system and a simple clustering algorithm, we outperform bag-of-words instance representations and compete with much more complex structured probabilistic models. Furthermore, we show that an oracle based on manually-labeled lexical substitutes yields yet substantially higher performance. Taken together, this provides evidence for a complementarity between word sense induction and lexical substitution that has not been given much consideration before.

1 Introduction

Lexical ambiguity, the phenomenon of words having multiple senses, is pervasive in language. Considerable attention in natural language processing (NLP) has been devoted to developing methods for describing lexical ambiguity and resolving it. Resolving lexical ambiguity has been argued to matter for many tasks from information retrieval (Stokoe, Oakes, and Tait 2003) and information extraction (Ciaramita and Altun 2006) to sentiment analysis (Wiebe and Mihalcea 2006) and machine translation (Carpuat and Wu 2007).

The most prominent approach to modeling and resolving lexical ambiguity is *word sense disambiguation* or WSD (Navigli 2009), where the task is to assign to each instance of a word a sense label from a fixed inventory such as WordNet (Fellbaum 1998). While useful, WSD methods suffer from two drawbacks. First, WSD typically uses supervised classification methods or lexical knowledge bases, which require large amounts of manually labeled data or large-coverage lexical resources. This makes WSD methods expensive to develop. Second, the assumption that a complete and appropriate sense inventory is available turns out to be problematic for various reasons, including resource coverage, domain-specific usages, or questions of sense granularity (Edmonds and Kilgarriff 2002; Snyder and Palmer 2004).

ID	Instance (word in context)	Cluster	Substitutes
1	... consistently reading **papers** with poor English...	1	*article*, *manuscript*
2	... while reading an item in the English **paper** today ...	2	*newspaper*, *periodical*
3	... **papers** may use previously published material...	1	*article*, *publication*
4	... the material uses fancy **paper**...	3	*pulp*, *parchment*

Table 1: Example of WSI clusters and lexical substitutions for four instances of the noun *paper*.

These problems have motivated research in *word sense induction* or WSI (Schütze 1998). WSI aims to induce word senses from unannotated corpora, typically by performing clustering. Unlike in WSD, word senses therefore arise dynamically from the corpora under consideration and can be tailored to the requirements of the scenario at hand. In this manner, WSI avoids the pitfalls of WSD outlined above. As an example for WSI, consider the (illustrative) clusters for some instances of *paper* shown in Table 1. Cluster 1 has the interpretation "paper as a scholarly article", cluster 2 corresponds to "paper as a newspaper", while cluster 3 corresponds to "paper as material", without making finer-grained distinctions. For comparison, WordNet lists seven senses for the noun *paper*, among which one for the scholarly article, three for the newspaper, and two for the material.

Importantly, the ability of WSI to identify meaningful word senses hinges on the extent to which the senses correlate with distinct contexts, as per the distributional hypothesis (Harris 1954). Thus, a major challenge for WSI is that many contexts are ambiguous or not discriminative enough, particularly at a bag-of-words level. This is also illustrated in Table 1: in terms of context words overlap, instances 1 and 2 as well as instances 3 and 4 are rather similar, rendering it difficult to cluster them into separate sense clusters.

Our paper specifically addresses the problem of noisy contexts. We start from the hypothesis that WSI can benefit from complementing contextual information with a different description of instance meaning, namely lexical substitutes. *Lexical substitution* or LS (McCarthy and Navigli 2007) is another relatively recent alternative to WSD that completely

avoids the notion of predefined word senses. Instead, it describes the meaning of each word instance in terms of a set of contextually appropriate substitutes (or paraphrases, comprising a *paraset*), as illustrated in the right-hand column of Table 1. We believe that context and lexical substitutes are complementary: context, as used in standard WSI approaches, provides a *syntagmatic* view of instance meaning by representing the target instance in terms of its context words. Lexical substitutes, in contrast, provide a syntagmatic as well as a *paradigmatic* view by representing the target instance in terms of contextually appropriate replacements that are nonetheless still similar to the target instance. In particular, we expect that LS can ameliorate the problem of ambiguous and non-discriminative contexts. This is illustrated in Table 1, where instances 1 and 3 share the substitute *article* (and therefore should end up in the same cluster), while instances 2 and 4 have a disjoint set of substitutes (and therefore should end up in different clusters). Indeed, previous analyses have found that lexical substitutes roughly reproduce word sense groupings while providing additional information (Kremer et al. 2014).

To the best of our knowledge, no previous study of WSI has investigated whether lexical substitutes can be leveraged for better WSI. This paper gives an affirmative answer to this question by making two contributions: (1) we demonstrate that WSI based on automatically predicted lexical substitutes outperforms a competitive context-based WSI model on the SEMEVAL-2010 benchmark (Manandhar et al. 2010) using a simple clustering approach; (2) we annotate a subset of that dataset with gold-standard lexical substitutes, finding even substantially higher performance. Thus, future improvements in LS modeling should translate to further improvements for WSI. Taken together, these results show a previously overlooked connection between WSI and LS which could translate into more robust joint models for both.

2 Related Work

A number of WSI approaches have been proposed in the literature. The standard approach is that of context clustering (Schütze 1998), in which the contexts of word instances are represented as individual second-order distributional vectors and then grouped into sense clusters (Jurgens and Stevens 2010). A related approach is to represent instances as nodes in a similarity graph which is clustered using graph-clustering algorithms (Korkontzelos and Manandhar 2010; Hope and Keller 2013b). The currently best-performing approaches rely on manually constructed probabilistic latent variable models (Van de Cruys and Apidianaki 2011; Lau et al. 2012; Bartunov et al. 2016; Komninos and Manandhar 2016), which structurally encode the dependencies between a word's sense and its context. Başkaya and Jurgens (2016) combine some of these models in an ensemble.

The other line of research relevant for our work is lexical substitution. LS was established as a paradigm with the SEMEVAL-2007 Lexical Substitution shared task (McCarthy and Navigli 2007). This task also provided the first LS dataset: a lexical sample dataset containing 200 instances of 20 words. A sample of this dataset was again used for the SEMEVAL-2010 Cross-Lingual Lexical Substitution task (McCarthy, Sinha, and Mihalcea 2013), where the substitutes were manually translated into Spanish. Another, considerably larger lexical sample is that of Biemann (2012), constructed through an iterative three-step crowdsourcing. The sample comprises 24,647 contexts for 1,012 frequent words, but is restricted to nouns only. In contrast to these two lexical sample datasets, the dataset presented by Kremer et al. (2014) is a crowdsourced all-words lexical substitution dataset, which comprises 15,629 words from 2,474 contexts.

The availability of these corpora boosted work on computational models for automatically predicting lexical substitutes. Nevertheless, apart from a few exceptions (Szarvas, Biemann, and Gurevych 2013; Hintz and Biemann 2015), most opted for unsupervised approaches. A common theme among LS models is to construct a contextualized distributional representation for a specific instance, e.g., by modifying the first-order distributional vector for a word so as to reflect the contribution of the context of the specific instance. The substitutes are then ranked by computing the similarity between their first-order vectors and the contextualized instance vector (Erk, McCarthy, and Gaylord 2013; Melamud, Levy, and Dagan 2015; Melamud, Dagan, and Goldberger 2015; Melamud, Goldberger, and Dagan 2016; Roller and Erk 2016). Abualhaija et al. (2017) frame lexical substitution as an optimization problem and solve it using metaheuristic approaches such as simulated annealing.

Our work combines WSI and LS. The study conceptually most similar to ours is (Başkaya et al. 2013), who used a language model to generate substitutes for target words and then constructed a distributional model over word-substitute pairs using S-CODE (Maron, Bienenstock, and James 2010). Those representations were then clustered to obtain a fixed number of sense clusters. Their work treats the substitutes merely as an intermediate representation and does not make the conceptual link to lexical substitutions. Indeed, being created by a language model, these substituted are not guaranteed to be paradigmatically related to the target words. Furthermore, their study is more limited in its experimental setup, in that they use the same number of senses for all words and do not evaluate the quality of context-based and substitute-based representations individually.

Another study that uses similar methods is (Cocos and Callison-Burch 2016), who induced word senses by clustering paraphrases from the Paraphrase Database PPDB (Pavlick et al. 2015), constructed using a bilingual pivoting method. Given two words that are potential paraphrases of each other (including lexical substitutes), they obtain their similarity score by either using precomputed paraphrase similarity score from PPDB (computed using a supervised regression model) or by comparing their second-order paraphrase scores (computed as similarity between two vectors of PPDB scores). Although the study frames its task as WSI, it does not evaluate on standard WSI datasets nor compares against state-of-the-art WSI methods. Our approach is also considerably simpler, requiring much less machinery.

3 Method

As argued in the introduction, our hypothesis is that lexical substitutes (the parasets) can improve on context-based

representations for clustering-based WSI. Thus, our method works by first computing the similarity of word instances based on parasets or a combination of parasets and contexts, and then performs clustering using a standard clustering algorithm, in our case affinity propagation.

3.1 Instance Similarity Measures

For each word separately, we construct a similarity matrix that stores the information about the similarity of each pair of instances. We experiment with three similarity measures:

CTX: Motivated by the robustness of additive models in computational semantics (Schütze 1998; Mitchell and Lapata 2010; Wieting et al. 2016), we construct the context-based representation of an instance by averaging the word embeddings for the content words found in the context (target word included). To ensure reproducibility, we use word embeddings trained on part of Google News dataset covering about 100 billion words (Mikolov et al. 2013).[1] The similarity of two word instances is then simply calculated as the positive cosine similarity of their instance embeddings. We use this similarity as the baseline.

AUTOLS: Following our hypothesis, this measure forgoes context words in favor of averaging the embeddings of the target word's lexical substitutes, as predicted by an automatic lexical substitution model (cf. Section 4). In case of multiword substitutes, we average the constituents' embeddings. We use the same word embeddings as for CTX, and again rely on positive cosine similarity.

AUTOLS+CTX: We combine the previous two measures by simply averaging their similarity predictions. This can be interpreted as "late fusion" (Kiela and Clark 2015).

In Section 6 we consider the setup with gold-standard (i.e., manually produced) lexical substitutes. For this setup, we define two additional instance similarity measures, GOLDLS and GOLDLS+CTX. These measures correspond to AUTOLS and AUTOLS+CTX, respectively, but use gold-standard instead of system-produced parasets.

3.2 Instance Clustering

We cluster the word instances using *affinity propagation* (Frey and Dueck 2007), an iterative clustering algorithm based on the concept of "message passing". The messages passed between the instances describe two values: evidence that instance k should be the exemplar for instance i (responsibility $r(i,k)$) and evidence that instance i should select the instance k to be its exemplar (availability $a(i,k)$). Formally, these quantities are given by:

$$\begin{aligned} r(i,k) &\leftarrow s(i,k) - \max_{\forall k' \neq k} \big(a(i,k') + s(i,k')\big) \\ a(i,k) &\leftarrow \min\big(0, r(k,k) + \sum_{\forall i' \notin \{i,k\}} r(i',k)\big) \end{aligned} \quad (1)$$

where $s(i,k)$ is a similarity score between instance i and k (matching the corresponding entry in a given similarity matrix). Specifically, the diagonal entries, $s(i,i)$, represent the input preference, indicating the likelihood of instance i becoming an exemplar. The algorithm starts by setting r and s to 0 for all the instances and then iterating until convergence. To avoid numerical oscillations, both $r(i,k)$ and $a(i,k)$ are linearly interpolated with their values from the previous iteration, controlled by a damping factor λ. We use the algorithm's default hyperparameters: factor λ of 0.5 and at most 200 iterations with convergence reached after 15 iterations with no change in the number of estimated clusters. We use the affinity propagation implementation of *scikit-learn* (Pedregosa et al. 2011).

The reason for using affinity propagation as our clustering algorithm is that it does not require the exact number of clusters to be set up front, a nuisance for many other clustering algorithms.[2] This is particularly important for WSI, where each word comes with different set of senses and whose number is difficult to establish automatically. However, in affinity propagation, the number of clusters is indirectly controlled by the value of the input preference. By adjusting the input preference, one can steer the algorithm toward producing more clusters (high input preference) or less clusters (low input preference), though eventually the algorithm will pick the final number of clusters to suit the data.

To investigate the effect of the number of clusters on WSI, we experiment with two settings for input preference: (1) the default value, set to the median of the similarity matrix, and (2) a value fine-tuned for each POS. In the latter case, for each POS and for each instance similarity measure, we perform a line search, minimizing the difference between the average predicted number of clusters and the average number of senses in WordNet for that POS. This effectively introduces *WordNet priors* on the number of clusters. The use of priors does not reduce the generality of our approach, as information about the average number of word senses across POS is typically available.

4 Experimental Setup

We carry out all of our experiments on the standard SEMEVAL-2010 WSI dataset (Manandhar et al. 2010), used for the shared task as well as for follow-up research on WSI. We will first present the results of our method in a "production mode", i.e., using system-produced lexical substitutes and following the official evaluation setup of the shared task (cf. Section 4.2). We will then provide a more detailed analysis on a 20-word dataset sample, giving insight into what performance gains could be obtained when using human-produced lexical substitutes (Section 6).

4.1 Data

The SEMEVAL-2010 WSI dataset is split into a training and test portion, and covers a sample of 50 nouns and 50 verbs. The training portion (879,807 contexts) was compiled in a semi-automatic manner, using WordNet 3.1 to automatically generate Yahoo! search queries, which were then manually checked by the task organizers. In contrast, the test portion

[1] https://code.google.com/archive/p/word2vec/

[2] There is a number of other algorithms that do not require the number of clusters to be set up front, such as Markov Clustering (Van Dongen 2000) and Chinese Whispers (Biemann 2006).

(8,915 contexts) was sampled from OntoNotes (Hovy et al. 2006), a large linguistic resource which, among other annotations, provides coarse-grained sense annotations of words in contexts. The average number of senses per word is 3.79. Since the training portion of the dataset is not sense-tagged and only serves as an additional source of word contexts, we use only the test portion in our experiments.

4.2 Evaluation

We follow the official evaluation setup and the official evaluation scripts, which comprises both supervised and unsupervised evaluation. For the supervised (i.e., WSD) evaluation, the induced sense labels are first converted to WordNet 3.1 labels via a mapping heuristic (Agirre and Soroa 2007) and then evaluated using (supervised) recall (SR). The mapping is learned on either 80% or 60% of the test set and amounts to determining which gold sense label appears most often with an induced sense label. To account for randomness, evaluation is repeated using five different mapping-test splits and averaging the obtained scores. We report results on the 80%-20% split. See (Manandhar et al. 2010) and (Agirre and Soroa 2007) for more details.

In contrast to the supervised evaluation, no mapping is performed for the unsupervised evaluation, where induced and gold sense labels (clusters) are directly compared using paired F-score (Artiles, Amigó, and Gonzalo 2009) and V-measure (Rosenberg and Hirschberg 2007). The paired F-score computes the harmonic mean of precision and recall by treating as true positives all instance pairs that are clustered together in both induced and gold sense clusters. The V-measure computes the harmonic mean of cluster homogeneity and cluster completeness.

As noted in (Manandhar et al. 2010), both V-measure and paired F-score exhibit biases regarding the number of clusters: V-measure is biased toward a higher number of clusters, whereas the paired F-score – despite averaging between pairwise precision and recall – penalizes a higher number of clusters. In addition, Pedersen (2010) notes that the V-measure is easily beat by random baselines. To account for these deficiencies, and following (Wang et al. 2015), for ease of comparison we also report the geometric mean of the two unsupervised measures, denoted by UAVG.

The SEMEVAL-2010 task organizers provided three baselines: *MFS* (most frequent sense – labels all test instances with the most frequent sense according to the mapping function), *1cl1inst* (one cluster per instance – gives each instance its own label),[3] and *Random* (gives all instances a random sense label). These baselines are complementary to each other and very competitive under some of the above metrics.

4.3 Preprocessing

We tokenize and lowercase the contexts using the *nltk* library (Bird, Klein, and Loper 2009) and keep only content words. We also preprocess the substitutes: we lemmatize them and then remove all the symbols except hyphen (-) and apostrophe ('). Additionally, we remove all multiword substitutes that contain the negation tokens (*not* or *n't*), to avoid the problem of modeling negation in distributional spaces (e.g., *beautiful* should be similar to *not ugly*).

4.4 Lexical Substitution Model

To obtain system-produced parasets, we use *context2vec* (Melamud, Goldberger, and Dagan 2016), one of the currently best-performing lexical substitution models. This model was proposed as a way of efficiently learning generic context embeddings, which could then be used across various natural language processing tasks, including lexical substitution. In a nutshell, *context2vec* builds upon the architecture of *word2vec*'s CBOW (Mikolov et al. 2013), but turns to a more expressive bidirectional LSTM architecture instead of simply averaging embeddings in a fixed window. For our experiments, we use a pre-trained model presented in (Melamud, Goldberger, and Dagan 2016) and simply take the top 15 lexical substitutes to serve as a system-produced paraset.

5 Word Sense Induction on SemEval-2010

The first experiment comprises the evaluation of our models on the complete SEMEVAL-2010 dataset, using official evaluation scripts provided by the organizers. The results are reported in Table 2. Within the supervised evaluation (SR score), our clustering model based on context similarity alone (CTX) performs very well, outperforming the baselines and all other competitor models, even the more complex ones, such as (Lau et al. 2012) and (Başkaya and Jurgens 2016). The performance improves further when using a model based on system-produced parasets (AUTOLS). Finally, combining the context-based and paraset-based similarity scores (AUTOLS+CTX) yields the overall best results. Using WordNet polysemy as the prior (+WN PRIOR models) results in a considerably lower number of clusters, and also reduces the SR scores.

Unsupervised evaluation gives a less clear picture due to the aforementioned biases of the V-measure metric and the paired F-score. Consequently, a baseline that uses one cluster per instance (*1cl1inst*) tops all other methods in terms of V-measure, whereas a baseline that labels all instances with a single sense (*MFS*) outperforms other methods in terms of paired F-score. The models of ours that use the default input preference for affinity propagation produce a larger number of clusters and hence perform much better under the V-measure, whereas the models with WordNet number of clusters produce a lower number of clusters and hence perform better under the paired F-score. In terms of the combined UAVG score, our models outperform all baselines, while the AUTOLS and AUTOLS+CTX models with WordNet priors perform on par with the state-state-of-the art model of (Korkontzelos and Manandhar 2010).

The overall tendency is for AUTOLS and AUTOLS+CTX models to consistently outperform their CTX counterparts. This trend also holds for both POS tags in the dataset. These results support our initial hypothesis that lexical substitutes complements context-based information, thus ameliorating problems of possibly noisy contexts.

[3] This baseline is not mentioned in the original paper, but was presented in the post-paper evaluation. See `https://www.cs.york.ac.uk/semeval2010_WSI/task_14_ranking.html`.

Model	SR (80%/20%)			V-measure			Paired F-score			UAVG	#C
	All	N	V	All	N	V	All	N	V	All	All
CTX	.693	.661	.740	.225	.263	.171	.196	.184	.215	.210	12.50
AUTOLS	.722	.668	.801	.268	.296	.227	.219	.209	.233	.242	13.79
AUTOLS+CTX	**.750**	**.707**	**.811**	**.248**	**.312**	**.232**	.220	.214	.230	.250	13.38
CTX (+WN prior)	.664	.614	.724	.145	.166	.114	.409	.431	.377	.243	2.94
AUTOLS (+WN prior)	.694	.644	.765	.195	.225	.151	.399	.413	.377	.279	3.59
AUTOLS+CTX (+WN prior)	.693	.641	.769	.191	.206	.168	**.423**	**.443**	**.393**	**.284**	3.25
HERMIT (Jurgens and Stevens 2010)	.583	.536	.653	**.162**	.167	.156	.267	.244	.301	.208	1.78
UoY (Korkontzelos and Manandhar 2010)	.624	.594	.668	.157	.206	.085	.498	.382	.666	**.280**	11.54
Duluth-WSI-SVD-Gap (Pedersen 2010)	.587	.532	.667	.000	.000	.000	**.633**	**.570**	**.724**	.000	1.02
NMF_{lib} (Van de Cruys and Apidianaki 2011)	.626	.573	.702	.118	.135	.094	.453	.422	.498	.231	4.80
NMF_{con} (Van de Cruys and Apidianaki 2011)	.603	.545	.688	.039	.039	.039	.602	.546	.684	.153	1.58
HDP+pos.+dep. (Lau et al. 2012)	**.680**	**.650**	**.720**	–	–	–	–	–	–	–	–
SNN_{swf} (Hope and Keller 2013a)	–	–	–	–	**.328**	**.246**	–	.144	.132	–	32.31
Ensemble* (Başkaya and Jurgens 2016)	**.680**	–	–	–	–	–	–	–	–	–	–
Most frequent sense (MFS)	**.587**	**.532**	**.666**	.000	.000	.000	**.635**	**.727**	**.570**	.000	1.00
One cluster per instance (1cl1inst)	.000	.000	.000	**.317**	**.256**	**.358**	.001	.001	.001	.018	89.15
Random	.573	.515	.657	.044	.046	.042	.319	.341	.304	**.118**	4.00

Table 2: Performance scores on the SEMEVAL-2010 WSI dataset (50 nouns and 50 verbs) of the proposed models (top two sections), models from the literature (middle section; results that are not publicly available are omitted), and the three baselines (bottom section). Best results in each group are shown in bold. Column #C shows the average number of obtained clusters. The model marked with * is not entirely comparable with ours as it uses a different sense mapping function.

6 Upper-Bound Performance

The experiment in Section 5 demonstrates that lexical substitutes obtained by a state-of-the-art LS system can improve WSI performance. A natural follow-up is to ask what the upper-bound performance for a lexical substitution-based WSI approach is, i.e., how much would WSI performance improve if ground-truth lexical substitutes were available. We investigate this question by repeating the first experiment on a sample of the SEMEVAL-2010 test portion for which we collect human-provided substitutes. We use the models with WordNet polysemy priors for these experiments.

6.1 Collecting Gold Substitutes

We selected 20 words (10 verbs and 10 nouns) from the SEMEVAL-2010 dataset and asked four annotators (students of English)[4] to provide substitutes for 50 instances of each word. The annotation task was set straightforwardly: the annotators were presented with a sentence containing a marked target word and were asked to provide as many substitutes as they deem appropriate (in any order). If appropriate, they were allowed to provide multiword substitutes. Conversely, in cases where the target word was a part of a multiword expression, the annotators were asked to provide substitutes for the whole phrase (most common cases are phrasal verbs, e.g., *stick out*). We did not require that the substitutes fit perfectly into the syntactic context, but instead allowed for small discrepancies due to differences in the use of determiners and prepositions. The annotators were forbidden to use any kind of lexical resource during the annotation.

To compile the final lexical substitute-annotated sample, we took the annotators' paraset union for each of the instances. We considered majority voting as well, but concluded that such approach is not as meaningful on such a small set of annotators – the best substitutes may be proposed only once and still be entirely reasonable. We make this dataset publicly available.[5]

6.2 Results

Table 3 shows the results of our models on this sample. We observe that the supervised evaluation scores are lower than that from Table 2. We ascribe this to the quality of mapping: as the mapping files were given only for a complete test portion of the dataset, we created them anew for our 20-word sample. Taking into account that our sample is far smaller in size than the original test portion, we ended up with a lower-quality mapping and in turn much lower results. In contrast, the unsupervised evaluation results are in line with those on a complete test portion.

The comparison between AUTOLS and GOLDLS shows that human-produced lexical substitutes further improve the performance of WSI models, as we expected. The gains are most evident in the terms of paired F-score and UAVG score even though the average numbers of clusters are similar. In terms of supervised recall and V-measure, the two model families perform similarly.

As a more qualitative analysis, consider Figure 1, which shows a two-dimensional representation of the instance spaces for the verb *commit* according to context similarity

[4] We ran a small preliminary study with both students of English and native BE speakers, and observed that students on average tend to generate more (appropriate) substitutes than the native speakers.

[5] http://takelab.fer.hr/lexsubclu/

Model	SR (80%/20%)			V-measure			Paired F-score			UAVG	#C
	All	N	V	All	N	V	All	N	V	All	All
CTX (+WN prior)	.207	.162	.210	.191	.193	.189	.431	.426	.436	.287	3.20
AUTOLS (+WN prior)	**.221**	**.176**	**.218**	.261	.249	.274	**.475**	.438	**.511**	.352	3.30
AUTOLS+CTX (+WN prior)	.199	.162	.206	**.318**	**.334**	**.303**	.445	**.442**	.448	**.376**	8.35
GOLDLS (+WN prior)	**.227**	**.180**	.244	**.313**	**.316**	.311	.536	.545	.528	.410	3.15
GOLDLS+CTX (+WN prior)	.215	.164	**.250**	.298	.223	**.372**	**.582**	**.548**	**.615**	**.416**	2.40

Table 3: Evaluation on a 20-word sample of SEMEVAL-2010 test portion with system- and human-produced lexical substitutes.

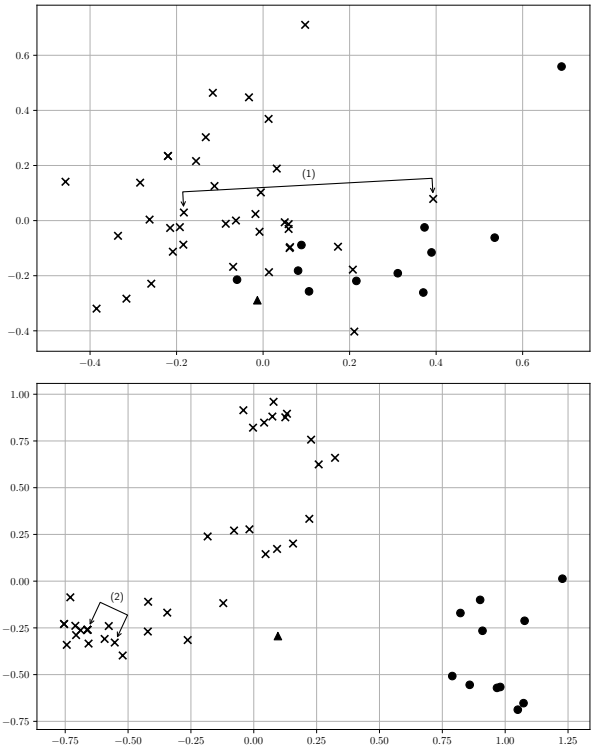

Figure 1: PCA transformation of context (upper) and paraset (lower) embedding space for instances of the word *commit*. Different senses of the word are denoted with different symbols. Example instances corresponding to the same sense are closer in the paraset space (2) than in context space (1).

(above) and gold paraset similarity (below). The paraset similarity separates the senses more clearly, in particular moving outliers further towards their sense centers. Two such instances of the first sense of *commit* ("perform an act") are shown in Table 4: since *commit* is used here as a light verb ("commit perjury, commit sin"), the lexical contexts have little in common. In contrast, the lexical substitutes bring out the shared meaning component of executing an action.

6.3 Mixing Predicted and Manual Substitutes

Our final question is how much better lexical substitution systems (and their substitutes) need to become for their performance to converge towards the ceiling performance we established above. To this end, we performed another evaluation on our 20-word dataset sample. We draw predictions from the best-performing model from Section 5, i.e., Table 2, namely AUTOLS+CTX (+WN prior), and the human-produced substitutes from Section 6. For each instance, we start off with an empty set and randomly fill it up with $r\%$ of human-produced substitutes and $(100 - r)\%$ of system-produced ones, with 100% corresponding to the cardinality of the complete human-produced paraset for that instance. To obtain robust performance estimates, we repeat this procedure 30 times and average the scores for each ratio $r \in \{0, 10, 20, \ldots, 100\}$.

The performance curves are shown in Figure 2. We observe a linear improvement in unsupervised WSI scores with the increase of human-produced substitute ratio. As the performance curves do not plateau out, we conclude that the lexical substitution models should improve quite a bit before we can rely solely on their substitutes for WSI.

Instance 1 (commit.v.40)
Well if, if he was using steroids during his testimony or prior to his testimony, obviously he has **committed** perjury and has tried to mislead the Congress in a – in an ongoing investigation.
Gold substitutes: *perpetrated, executed, exerted, done, performed, carried out*
Instance 2 (commit.v.57)
Yet some people are advancing a chilling casuistry: that what we are seeing is somehow the understandable result of the historical sins **committed** by the Turks in the 16th century.
Gold substitutes: *carried out, conducted, made, done, performed*

Table 4: A pair of instances for WordNet sense 1 ("perform an act") of the verb *commit*. These instances correspond to the ones on Figure 1.

7 Discussion and Conclusion

Lexical substitution and word sense induction offer two perspective on word senses that can complement each other – not only conceptually, but, according to our results, also empirically. Still, these perspectives are rarely brought together in the literature. In this paper, we presented a clustering-based model for word sense induction that follows this intuition and leverages the information found in lexical substitutes, even if they are noisy substitutes predicted by

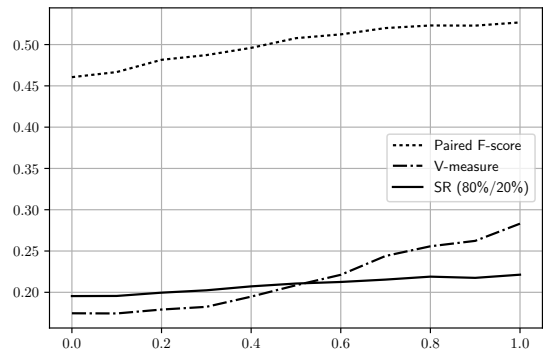

Figure 2: Model performance for different ratios of human-provided and automatically predicted lexical substitutes (1.0 = GOLDLS+CTX (+WN prior)).

automatic methods. In our experiments on the standard SEMEVAL-2010 dataset, we found that our method can considerably outperform models that consider only the context of the occurrences to be clustered, using an extremely simple clustering approach. Further, both kinds of representations can be combined for even better results. Evidently, computational models of lexical substitution pick up information that it not easily accessible in a standard bag-of-words representation. Whether this effect is more akin to context feature selection or to a smoothing effect brought about by the substitutes remains to be explored.

We also explored the trade-off between automatically predicted and human-provided substitutes for word sense induction on a small manually annotated dataset, which we make freely available. We established that the ceiling performance for using manual substitutes is still considerably higher, which raises the prospect that future improvements in lexical substitution models could also further improve word sense induction.

In this study, we have only used lexical substitutes in word sense induction, but not vice versa. In future work, we will explore whether the clusters found by word sense induction can also help lexical substitution systems, for example by giving them a prior that two instances should or should not be assigned similar substitutes. Further along these lines, we would expect that a joint model of word sense induction and lexical substitution should be able to bring both conceptual perspectives together and further increase the quality and precision of word sense modeling, a longstanding desideratum in the natural language processing community.

8 Acknowledgements

We would like to thank the four annotators that annotated our data. We would also like to thank the reviewers for their insightful comments. This work has been fully supported by the Croatian Science Foundation under the project UIP-2014-09-7312. Sebastian Padó acknowledges support by the DFG through SFB 732 (project D10).

References

Abualhaija, S.; Miller, T.; Eckle-Kohler, J.; Gurevych, I.; and Zimmermann, K.-H. 2017. Metaheuristic approaches to lexical substitution and simplification. In *Proceedings of EACL*, 870–880.

Agirre, E., and Soroa, A. 2007. SemEval-2007 Task 2: Evaluating word sense induction and discrimination systems. In *Proceedings of SemEval*, 7–12.

Artiles, J.; Amigó, E.; and Gonzalo, J. 2009. The role of named entities in web people search. In *Proceedings of EMNLP*, 534–542.

Bartunov, S.; Kondrashkin, D.; Osokin, A.; and Vetrov, D. 2016. Breaking sticks and ambiguities with adaptive skip-gram. In *Artificial Intelligence and Statistics*, 130–138.

Başkaya, O., and Jurgens, D. 2016. Semi-supervised learning with induced word senses for state of the art word sense disambiguation. *Journal of Artificial Intelligence Research* 55:1025–1058.

Başkaya, O.; Sert, E.; Cirik, V.; and Yuret, D. 2013. AI-KU: Using substitute vectors and co-occurrence modeling for word sense induction and disambiguation. In *Proceedings of SemEval*, 300–306.

Biemann, C. 2006. Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraph*, 73–80.

Biemann, C. 2012. Turk bootstrap word sense inventory 2.0: A large-scale resource for lexical substitution. In *Proceedings of LREC*, 4038–4042.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly.

Carpuat, M., and Wu, D. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of EMNLP-CoNLL*, 61–72.

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

Cocos, A., and Callison-Burch, C. 2016. Clustering paraphrases by word sense. In *Proceedings of NAACL-HLT*, 1463–1472.

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

Erk, K.; McCarthy, D.; and Gaylord, N. 2013. Measuring word meaning in context. *Computational Linguistics* 39(3):511–554.

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

Frey, B. J., and Dueck, D. 2007. Clustering by passing messages between data points. *Science* 315:972–976.

Harris, Z. S. 1954. Distributional structure. *Word* 10(2–3):146–162.

Hintz, G., and Biemann, C. 2015. Delexicalized supervised German lexical substitution. In *Proceedings of the GermEval GSCL Workshop*, 11–16.

Hope, D., and Keller, B. 2013a. MaxMax: A graph-based soft clustering algorithm applied to word sense induction. In *Proceedings of CicLing*, 368–381.

Hope, D., and Keller, B. 2013b. UoS: A graph-based system for graded word sense induction. In *Proceedings of SemEval*, 689–694.

Hovy, E.; Marcus, M.; Palmer, M.; Ramshaw, L.; and Weischedel, R. 2006. OntoNotes: The 90% solution. In *Proceedings of NAACL*, 57–60.

Jurgens, D., and Stevens, K. 2010. HERMIT: Flexible clustering for the SemEval-2 WSI task. In *Proceedings of SemEval*, 359–362.

Kiela, D., and Clark, S. 2015. Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of EMNLP*, 2461–2470.

Komninos, A., and Manandhar, S. 2016. Structured generative models of continuous features for word sense induction. In *Proceedings of COLING*, 3577–3587.

Korkontzelos, I., and Manandhar, S. 2010. UoY: Graphs of unambiguous vertices for word sense induction and disambiguation. In *Proceedings of SemEval*, 355–358.

Kremer, G.; Erk, K.; Padó, S.; and Thater, S. 2014. What substitutes tell us – analysis of an "all-words" lexical substitution corpus. In *Proceedings of EACL*, 540–549.

Lau, J. H.; Cook, P.; McCarthy, D.; Newman, D.; and Baldwin, T. 2012. Word sense induction for novel sense detection. In *Proceedings of EACL*, 591–601.

Manandhar, S.; Klapaftis, I. P.; Dligach, D.; and Pradhan, S. S. 2010. SemEval-2010 Task 14: Word sense induction & disambiguation. In *Proceedings of SemEval*, 63–68.

Maron, Y.; Bienenstock, E.; and James, M. 2010. Sphere embedding: An application to part-of-speech induction. In *Proceedings of NIPS*, 1567–1575.

McCarthy, D., and Navigli, R. 2007. SemEval-2007 Task 10: English lexical substitution task. In *Proceedings of SemEval*, 48–53.

McCarthy, D.; Sinha, R.; and Mihalcea, R. 2013. The cross-lingual lexical substitution task. *Language Resources and Evaluation* 47(3):607– 638.

Melamud, O.; Dagan, I.; and Goldberger, J. 2015. Modeling word meaning in context with substitute vectors. In *Proceedings of NAACL-HLT*, 472–482.

Melamud, O.; Goldberger, J.; and Dagan, I. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of CoNLL*, 51–61.

Melamud, O.; Levy, O.; and Dagan, I. 2015. A simple word embedding model for lexical substitution. In *Proceedings of the NAACL VSM-NLP Workshop*, 1–7.

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*, 3111–3119.

Mitchell, J., and Lapata, M. 2010. Composition in distributional models of semantics. *Cognitive Science* 34(8):1388–1429.

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

Pavlick, E.; Rastogi, P.; Ganitkevitch, J.; Van Durme, B.; and Callison-Burch, C. 2015. PPPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of ACL-IJCNLP*, 425–430.

Pedersen, T. 2010. Duluth-WSI: SenseClusters applied to the sense induction task of SemEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 363–366.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Roller, S., and Erk, K. 2016. PIC a different word: A simple model for lexical substitution in context. In *Proceedings of NAACL*, 1121–1126.

Rosenberg, A., and Hirschberg, J. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of EMNLP-CoNLL*, 410–420.

Schütze, H. 1998. Automatic Word Sense Discrimination. *Computational Linguistics* 24(1):97–124.

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

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

Szarvas, G.; Biemann, C.; and Gurevych, I. 2013. Supervised all-words lexical substitution using delexicalized features. In *Proceedings of HLT-NAACL*, 1131–1141.

Van de Cruys, T., and Apidianaki, M. 2011. Latent semantic word sense induction and disambiguation. In *Proceedings of ACL-HLT*, 1476–1485.

Van Dongen, S. M. 2000. *Graph Clustering by Flow Simulation*. Ph.D. Dissertation, Dutch National Research Institute for Mathematics and Computer Science, University of Utrecht.

Wang, J.; Bansal, M.; Gimpel, K.; Ziebart, B. D.; and Clement, T. Y. 2015. A sense-topic model for word sense induction with unsupervised data enrichment. *Transactions of the Association for Computational Linguistics* 3:59–71.

Wiebe, J., and Mihalcea, R. 2006. Word sense and subjectivity. In *Proceedings of COLING-ACL*, 1065–1072.

Wieting, J.; Bansal, M.; Gimpel, K.; and Livescu, K. 2016. Towards universal paraphrastic sentence embeddings. *Proceedings of ICLR*.