Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/taut20

Representing word meaning in context via lexical substitutes

Domagoj Alagić & Jan Šnajder

AUTOMATIKA
2021, VOL. 62, NO. 2, 239–248
https://doi.org/10.1080/00051144.2021.1928437

REGULAR PAPER

OPEN ACCESS Check for updates

Representing word meaning in context via lexical substitutes

Domagoj Alagić and Jan Šnajder

Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia

ABSTRACT

Representing the meaning of individual words is crucial for most natural language processing (NLP) tasks. This, however, is a challenge because word meaning often depends on the context. Recent approaches to representing word meaning in context rely on *lexical substitution* (LS), where a word is represented with a set of meaning-preserving substitutes. While face valid, it is not clear to what extent substitute-based representation corresponds to the more established sense-based representation required for many NLP tasks. We present an empirical study that addresses this question by quantifying the correspondence between substitute- and sense-based meaning representations. We compile a high-quality dataset annotated with lexical substitutes and sense labels from two well-established sense inventories, and conduct a correlation analysis using a number of substitute-based similarity measures. Furthermore, as recent work has demonstrated the efficacy of system-produced substitutes for word meaning representation, we compare human- and system-produced substitutes to determine the performance gap between the two. Lastly, we investigate to what extent the results translate to the fundamental semantic task of word sense induction (WSI). Our experiments show the validity of LS for word meaning in context representation and justify the use of system-produced substitutes for WSI.

ARTICLE HISTORY
Received 4 September 2020
Accepted 30 March 2021

KEYWORDS
Natural language processing; machine learning; lexical substitution; word meaning in context; word sense induction

1. Introduction

Natural language processing (NLP) deals with the understanding of language meaning. Almost all NLP tasks of practical interest, ranging from information retrieval and machine translation to question answering and chatbots, hinge on the ability to recognize the meaning of *words*. The task is aggravated by the fact that words can have different meanings depending on the context in which they are used. NLP has traditionally relied on two approaches to word meaning [1]: *relational semantics* and *distributional semantics*. The former uses the notion of a *sense* and defines word meaning by sense relations (synonymy, hypernymy, hyponymy, etc.) that the word bears to other words. For instance, one sense of "road" is synonymous to "route", while the other sense is synonymous with "means". The practical application of this idea is WordNet [2,3], a lexical database for English organized according to sense relations, which has been used for numerous NLP tasks.

Distributional semantics [4] also adopts the relational view, but, instead of sense relations, it considers the relations between words co-occurring in sentences. The principle is best summarized by the *distributional hypothesis* [5], which posits that the meaning of a word can be deduced by observing the word's contexts. For instance, one sense of the word "road" is defined by it occurring together with words "car" and "traffic". The distributional approach is the cornerstone of recent neural approaches to NLP [6], which gave rise to extensive research on pretrained language models for contextualized word representations (e.g. [7]).

Recently, a third approach to word meaning, especially apt for representing word meaning in context, has received increased attention in NLP: *lexical substitution* [8]. A lexical substitute is a meaning-preserving replacement for a word in context. For example, certain contexts warrant the substitution of "road" with "street", while for others "way" would be more suitable. The meaning of the word in context is then taken to correspond to the set of its lexical substitutes, also known as a *paraset* (*paraphrase set*). The task of automatically producing lexical substitutes has attracted considerable attention in NLP (e.g. [9,10]), and various models have been proposed that rely on system-produced substitutes for word meaning representation (e.g. [11,12]).

While lexical substitution (LS) intuitively appears to be a sensible approach to representing word meaning in context, it is by no means evident how it relates to sense-based representation. However, determining the correspondence between substitute- and sense-based meaning is important for at least two reasons. Firstly, many practical NLP applications require, for a given word in context, to explicitly identify its sense from a sense inventory such as WordNet, as in the *word sense disambiguation* [13] task, or to group together contexts pertaining to the same sense, as in the *word*

CONTACT Domagoj Alagić domagoj.alagic@fer.hr Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, Zagreb 10000, Croatia

sense induction (WSI) [14] task. Secondly, even when detecting senses is not an end goal in itself, it is important to have a way of validating substitution-based representations, which can be achieved by comparing it to the more established sense-based representation.

Our work aims to fill the above-mentioned gap by investigating the viability of using lexical substitutes for representing word meaning in context. We present an empirical study that quantifies the correspondence between substitute- and sense-based meaning representations. To this end, we compile a high-quality lexical sample dataset in English, with human-produced lexical substitutes and sense labels from two well-established sense inventories, WordNet 2020 [3] and OntoNotes [15].

Furthermore, as recent work has demonstrated the efficacy of system-produced lexical substitutes for word meaning representation, we directly compare human- and system-produced lexical substitutes to determine the performance gap between the two. Lastly, we investigate to what extent the results translate to a typical semantic task, namely WSI, where we consider simple WSI models and a state-of-the-art WSI model based on substitutes produced by a neural language model [12]. More concretely, our study addresses the following research questions:

- RQ1: To what extent does word meaning representation based on lexical substitutes correspond to sense-based meaning representation?
- RQ2: Does the correspondence significantly deteriorate if system-produced substitutes are used in lieu of human-produced ones?
- RQ3: Is there a difference between human- and system-produced substitutes when used for the WSI task?

The three research questions define a roadmap for vindicating the use of lexical substitutes in NLP systems. The departing question, addressed by RQ1, is the very plausibility of using lexical substitutes as a means for representing contextualized word meaning. Assuming this question is answered in the affirmative, and acknowledging the high costs of human-produced lexical substitutes, the next question, duly addressed by RQ2, considers whether automatically-produced substitutes are fit for the same job. Lastly, since NLP is about solving real-life tasks, findings eventually have to be validated on external benchmark tasks to be considered practically relevant, which is what RQ3 addresses using WSI as the prototypical lexicosemantic NLP task.

We argue that the above questions are crucial for validating the use of LS in NLP. To the best of our knowledge, our study is the first to address these questions. A secondary contribution of our work is the first dataset in English annotated with both sense annotations (single- and multi-sense) and lexical substitutes, where the latter are collected using a robust three-step annotation procedure. We make this dataset publicly available in hope of fostering further research on this topic.

2. Related work

Our study relates to two strands of NLP research: lexical substitution (LS) and word sense induction (WSI). We next review the most prominent work from these two tasks.

2.1. *Lexical substitution*

LS was first introduced at the SemEval-2007 shared task [8] and since attracted mostly unsupervised machine learning approaches, with an exception of a few feature-based supervised models [16,17]. The early unsupervised models computed similarities between the distributional representations of the context, target word, and the candidates to produce a ranking of plausible candidates [18,19]. The more recent models rely on pre-trained neural language models [9,10]. For training and evaluation of LS systems, a number of standard datasets have been compiled. Besides the dataset compiled for SemEval-2007 [8], the most popular ones are CoInCo [20] and TWSI 2.0 [21], covering a sizable number of words annotated through crowdsourcing efforts.

Another line of research focuses on the use of LS for semantic modelling. In [11], LS was used to build substitute-based word representations, which were then tested on the semantic similarity task and a series of extrinsic benchmarks, including dependency parsing and sentiment analysis. In contrast, the study in [22] used LS as a testbed for evaluating other representations, more specifically, compositional distributional semantic models.

While there exists ample work on LS systems and their applications, only a handful of studies addressed the key question of how LS corresponds to sense-based word meaning. The study in [23] used lexical substitutes as one of the tools to investigate how well words can be partitioned into senses. The results indicate that partitionability can be quantified quite well with intra-clustering clusterability measures based on lexical substitutes. Similarly, the study in [20] compared lexical substitutes to WordNet synsets (synonym sets) by devising a heuristic mapping between the two. The study showed that, while lexical substitutes correspond rather well to WordNet senses, they often induce subtle sense distinctions not covered by WordNet. Our study also investigates the relation between substitutes and senses but, instead of relying on heuristic mapping, we quantify the correspondence between LS and sense-based meaning representation by directly comparing the LS and sense-based similarity measures. We also explore to what extent that correspondence translates to the WSI task.

2.2. *Word sense induction*

WSI is a well-established NLP task that attracted many different approaches [13]. The early approaches used *context clustering*, which involves computing the distributional vectors of the observed word's contexts and grouping them into a number of sense clusters [14,24,25]. Subsequent approaches used *clustering word co-occurrence graphs* [26] and *probabilistic clustering*, in which word sense induction is formalized as a generative model [27,28].

Since the WSI task requires the representation of word meaning in context, it is a natural candidate for using LS for meaning representation. Indeed, the currently best-performing WSI approaches rely on LS as a means of capturing word meaning. The first such approach [29] included a simple 4-gram language model to generate lexical substitutes, which were then used to build a distributional model over word-substitute pairs. The more recent works opted for specialized LS models over standard language models, and also incorporated substitutes in a more straightforward manner. For instance, the approach in [30] relied on separate context and paraset representations (i.e. their similarities) to measure how similar the individual instances are. Apart from using the then-current state-of-the-art LS model, they also experimented with using human-produced lexical substitutes and showed that this leads to considerable performance improvements. The current state-of-the-art WSI model, proposed in [12], uses context only as an input for the pretrained BERT language model [7]. Their WSI model then samples substitutes from the vocabulary, represents them as bag-of-words vectors, and clusters them using hierarchical agglomerative clustering (HAC). In our study, we evaluate the model of [12] and compare it against a similar model with human-produced substitutes to determine the performance gap between the two.

When it comes to data, WSI models are almost exclusively evaluated on datasets from two shared tasks, SemEval 2010 [31] and SemEval 2013 [32]. The former is based around data labelled with single-sense annotations, while the latter uses multi-sense data, which is also reflected in the choice of evaluation metrics. Work in [30] complemented a subset of SemEval 2010 data with additional LS annotations. We use the dataset of [30] as a starting point for our study, but expand it with (multi-)sense annotations from a different sense inventory and refine the original lexical substitutes for quality. To the best of our knowledge, there is no other dataset for the English language annotated with both lexical substitutes and (multi-)sense word labels.

Table 1. Datasets summary (ON = OntoNotes, WN = English WordNet 2020).

	# Targets	# Contexts	Substitutes	Senses	Multisense
SemEval 2010 [31]	100	8915	✗	ON	✗
Alagić et al. (2018) [30]	20	1000	✓	ON	✗
This work	20	837	✓	ON & WN	✓ (WN)

3. Dataset annotation

Our study uses the dataset from [30] as a starting point, which in turn is a subset of SemEval-2010 dataset [15]. The SemEval dataset covers 50 verbs and 50 nouns across 8,915 contexts sampled from OntoNotes [15], a large linguistic resource with several annotation layers, including word senses. These were collected using an iterative process of refining the sense inventory and re-annotating the instances until a satisfactory inter-annotator agreement was reached. The study in [30] used a subset of this dataset consisting of 20 words (10 nouns and 10 verbs) across 1000 contexts, each additionally annotated with lexical substitutes. While we could in principle use this dataset to answer our research questions, as it is annotated with both lexical substitutes and word senses, it nonetheless suffers from two serious shortcomings that could threaten the validity of our study: (1) it is restricted to a single-sense (and arguably less standard) sense inventory (OntoNotes) and (2) the quality of lexical substitutes is manifestly low. To address this, we expanded and improved on both annotation layers: we additionally annotated word senses using the newly-released English WordNet 2020 [3] and we revised all lexical substitutes. The subsequent sections detail the two annotation efforts. Table 1 summarizes the three datasets and the differences between them, while Table 2 shows examples from our dataset.

3.1. *Complementary annotation of word senses*

While OntoNotes annotations provided in the SemEval-2020 dataset are unquestionably of high quality, all sense inventories introduce certain biases with respect to sense definitions and granularity. Thus, to improve the validity of our study, we decided to consider two complementary sense inventories: besides using the OntoNotes labels, we annotated all instances with sense labels from the recently-introduced English WordNet 2020 [3]. Unlike OntoNotes, we opted for multi-sense annotations, i.e. allowing one word in context to be labelled with a number of senses. Considering that no sense inventory can ensure contexts of completely disjoint senses, we believe that allowing multi-sense labels could bring us potentially valuable insights into the problem of capturing word meaning in context. Another side benefit of having WordNet annotations

Table 2. A few instances from our dataset, together with their WordNet 2020 sense labels and refined lexical substitutes.

road.n.132	
Sentence	*The convoy was travelling near the West Bank city of Ramallah on a stretch of* **road** *controlled by the Israeli military.*
Word sense(s)	an open way (generally public) for travel or transportation
Substitutes	*roadway* (4), *land* (2), *street* (2), *territory* (2), *section* (1), *lane* (1), *path* (1), *ground* (1), *way* (1), *route* (1)
road.n.125	
Sentence	*The fact that as we continue on this* **road** *and continue to talk about what's going to happen in Florida, the judges have really already spoken on this, the ultimate judges, the people.*
Word sense(s)	a way or means to achieve something
Substitutes	*path* (5), *direction* (4), *course* (4), *way* (2)
relax.v.3	
Sentence	*Secondly, I'm directing the Secretary to* **relax** *sanctions on American countries and citizens conducting business in Iraq that contributed to humanitarian reconstruction.*
Word sense(s)	make less severe or strict
Substitutes	*ease* (4), *loosen up* (4), *relieve* (3), *lessen* (3), *weaken* (3), *loosen* (2), *reduce* (2), *moderate* (1), *cut* (1), *minimize* (1), *lower* (1)
relax.v.21	
Sentence	*He was no longer able to* **relax** *in the presence of his parents and found it difficult to keep up a conversation with his mother or father, no matter the subject.*
Word sense(s)	1. become less tense, rest, or take one's ease 2. become less tense, less formal, or less restrained, and assume a friendlier manner
Substitutes	*be at ease* (5), *loosen up* (3), *be at peace* (3), *unwind* (1), *chill out* (1), *feel welcome* (1)

Note: The number next to a particular substitute denotes how many annotators kept it during the annotation.

is that, in contrast to OntoNotes and its sense inventory, WordNet is freely available and more widely used, which increases the practical value of our results.

To obtain the annotations, we asked five near-native speakers of English to label each instance (a target word in context) with appropriate WordNet 2020 senses. We allowed the annotators to select more than one sense if they found it necessary, e.g. in case of overly ambiguous contexts. In case they deemed no sense appropriate for a given context, they were asked to select the "None of the above" (NOTA) option.

The complete sense annotation took 38 person-hours. On average, the annotators selected multiple senses for 75 instances (9% of total instances) and selected NOTA for 32 instances (3.8%). Taking into account the low number of multi-sense annotations, we calculated the inter-annotator agreement only on the single-sense annotations, thus avoiding the notorious issue of calculating agreement on multilabel annotations. We used the Cohen's κ averaged over all ten annotator pairs: the observed agreement is 0.61, which is considered a substantial agreement [33].

To obtain the final WordNet sense labels for each of the instances, we decided to adopt two strategies, each yielding a different dataset variant:

- WN-Single – The word sense label is obtained via majority voting. In case of a tie, the instance is dropped;
- WN-Multi – Only senses chosen by the majority (i.e. three or more) of annotators are included in the final set of word senses of an instance. If none of the senses passed that threshold, the instance is dropped.

Table 2 shows sense annotations (represented by WordNet 2020 glosses) for a few instances from our dataset.

3.2. Revising lexical substitutes

Our manual inspection of the lexical substitutes dataset from [30] revealed deficiencies in the quality and consistency of annotations. More concretely, for many instances some of the substitutes provided by the annotators are arguably not preserving the meaning of the target word. What is more, in some cases the annotators were not consistent and provided substitutes not for the target word itself but for a surrounding sequence of words encompassing the target word, typically when the target word was part of a multiword expression (e.g. the target "road" in expression "down the road").

As the above issues would jeopardize the validity of our study, we decided to thoroughly revise the annotations. The revision was carried out in two steps. In the first step, we (the authors) manually inspected all substitute annotations across all instances aggregated across the five annotators, and identified the cases where one of the substitutes pertained to a multiword expression containing the target word. For instances in which all substitutes pertained to an expression, we revised the target word to be that particular expression if it was a semantically opaque expression (e.g. a phrasal verb), otherwise we discarded the instance. Conversely, for instances in which only some substitutes pertained to the expression, while others pertained to the individual target word, we removed only the former substitutes. In this step, we also corrected the spelling errors in the substitutes and lemmatized all substitutes. After this step, we ended up with 837 instances.

In the second revision step, we asked five annotators to go through all lexical substitutes and, without providing any of their own, discard the ones that they think do not preserve sentence meaning to a high degree or with which the sentence does not sound natural (while ignoring minimal syntactic alternations of context, e.g. differences in prepositions or articles). To obtain the revised lexical substitute sets (parasets) for an instance, we simply took all the substitutes that any of the five annotators decided to keep (i.e. we took the substitute union). This resulted in $\sim 20\%$ fewer lexical substitutes, indicating that this step was justified. In the end, this step took 56 person-hours.

With the two revision steps following the original annotation, the annotation can conceptually be conceived as consisting of three steps: (1) *elicitation*, in which annotators are asked to provide as many substitutes as they can think of, with the aim of not missing any relevant substitute, (2) *clean-up*, where experts manually revise the substitutes for consistency, and (3) *filtering*, where substitutes are checked again by the annotators and the low-quality substitutes are removed. We argue that this three-step procedure is optimal in the sense that it ensures high quality while at the same time preserving high coverage.

4. Word meaning in context via lexical substitutes

As argued in the introduction, the use of LS for representing word meaning in context raises the fundamental question of how this representation corresponds to word senses (RQ1). While it is clear that lexical substitutes do represent certain aspects of a word's meaning in context, the assumption is that it is in particular the sense-based meaning that is well represented. This assumption can be readily verified by comparing the contexts featuring the same target word: if a word's meaning in context indeed corresponds to senses, then the parasets should be identical for same-sense target words. In practice, however, due to granularity mismatch, as well as nuanced differences in meaning observed by Kremer et al. [20], we intuitively expect the sets of substitutes for same-sense words not to be perfectly identical but rather highly similar. In virtue of that, the extent to which substitute-based representation corresponds to sense-based representation can be taken to be equivalent to the degree of similarity between sets of lexical substitutes that correspond for same-sense words.

The experimental design of RQ1 is an operationalization of the above observation: we quantify the extent to which substitute-based representation correspond to sense-based representation by correlation analysis between similarities of parasets and sense matches. Furthermore, since similarities between parasets can be computed in a number of ways, we consider a number of standard similarity measures.

We then investigate how much the correspondence between substitute- and sense-based representation deteriorates if one switches to system-produced substitutes (RQ2). We follow the same experimental design as for RQ1, but use lexical substitutes predicted by a strong neural language model instead of substitutes provided by the annotators. From an NLP perspective, system-produced substitutes are easy to obtain, whereas human-produced substitutes constitute an ideal but unrealistic setup. The difference in correlation between the two determines the performance gap between word meaning representation based on system- and human-produced lexical substitutes. If this gap is not too large, using automatic LS systems for word meaning representation is a viable option.

4.1. Similarity measure correlation analysis

The correlation analysis was carried out on pairs of instances from our dataset. First, we generated all possible pairs from all the instances (leaving out symmetric and reflexive pairs). After that, for each instance pair, we checked whether their word sense labels match (either WN-Single or WN-Multi), which resulted in a $\binom{N}{2}$-sized binary similarity vector. We then repeated this process by using a similarity function that operates on pairs of parasets. Lastly, we computed the point-biserial correlation coefficient r_{pb} between the two vectors.[1] The higher the value of the correlation coefficient, the more the similarity between substitutes corresponds to matches between senses, and, consequently, the more substitute-based meaning representation corresponds to sense-based meaning representation.

We experiment with a number of different paraset-based similarity measures, including both standard methods as well as some methods proposed in LS research:

- ParaExact – 1 if both instances have the exact same paraset, 0 otherwise;
- ParaDice – Dice coefficient between the instances' parasets;
- ParaJaccard – Jaccard coefficient between the instances' parasets;
- ParaGAP – Generalized Average Precision (GAP) [34] between the instances' parasets. As that the score depends on which out of the two parasets serves as the reference, we compute GAP for both cases and take the average;
- ParaCosine – Cosine similarity between parasets encoded as score-based bag-of-words vectors over the substitute vocabulary;
- ParaCosineBin – Cosine similarity between parasets encoded as binary bag-of-words vectors over the substitute vocabulary;
- SenseExact – 1 if both instances are labelled with the same word sense(s), 0 otherwise. Other measures are compared against this one.

Above, "score" denotes a substitute's annotation frequency, i.e. how many annotators have provided that particular substitute for a given target word. In the case of ParaCosineBin, this score is replaced with a binary variable. The results of the correlation analysis are shown in the left half of Table 3. We show correlations for the OntoNotes inventory and for WordNet inventory (for both WN-Single and WN-Multi). Correlation is the lowest for the most rigid measure,

Table 3. Correlation between SenseExact and other paraset-based similarity measures (point-biserial correlation coefficient r_{pb}) for human- and system-produced lexical substitutes.

	Human-produced			System-produced		
Measure	ON	Single	Multi	ON	Single	Multi
ParaExact	0.056	0.062	0.064	–	–	–
ParaDice	**0.494**	0.531	0.523	**0.496**	0.465	**0.451**
ParaJaccard	0.475	0.517	0.510	0.493	0.464	0.449
ParaGAP	0.447	0.508	0.504	0.464	0.387	0.379
ParaCosine	0.481	**0.537**	**0.530**	0.403	0.342	0.339
ParaCosineBin	0.492	0.530	0.522	**0.496**	**0.466**	**0.451**

Note: All correlations are statistically significant with $p < 0.0001$. The highest correlation values for each setup are shown in bold.

ParaExact. This confirms our intuition, and the findings of [20], that lexical substitutes capture nuanced differences in meaning that escape sense distinctions, rendering unlikely a perfect match between sets of substitutes for same-sense words. On the other hand, the situation with other measures is not as clear: most correlation scores are quite close to each other and it is difficult to pinpoint a clear winner. Interestingly, the simplest measures, ParaDice and ParaJaccard, perform rather well. As for the sense inventories, WordNet seems more suitable for paraset-based similarity measures than OntoNotes, which is evident by the somewhat larger correlation scores. When comparing single-sense and multi-sense labels, we observe no major difference in correlation scores, which is expected since only a small fraction of instances was annotated with multi-sense labels (cf. Section 3.1). In conclusion, we observe substantial positive correlation for all considered paraset-based similarity measures, with the effect size depending on the choice of sense inventory.

4.2. *System-produced lexical substitutes*

To obtain system-produced substitutes, we adopt the LSDP system of [35], used in the state-of-the-art WSI model of [12]. The LSDP (*LS with Dynamic Patterns*) uses a language model to generate lexical substitutes by predicting the words that could replace the original target word. However, as simply doing so would allow the model to produce words that preserve merely the syntax but not the meaning, the authors introduced a clever trick, dubbed *dynamic patterns*, to alter the original context and make it more semantically constrained. For example, predicting the substitutes for target word "*brown*" in "*My dogs are brown*" by feeding the model "*My dogs are* ___" could result in words such as "*barking*", "*beautiful*", or "*outside*". Using a pattern, however, the context will be altered to "*My dogs are brown (or even*___*)*", which will hopefully steer the model into producing meaning-preserving substitutes of the target word. Since the language model may produce substitutes in inflected forms, an additional lemmatization step is applied to all substitutes.

For our experiments, we build the final instance parasets by simply taking the top 200 words produced by the BERT language model used in [12].[2] BERT is a powerful language model built on top of the recently proposed attention-only neural network architecture [36] and pretrained on large corpora. Considering its superb performance across many NLP tasks, its use for LS is very well justified.

To answer RQ2, we repeated the correlation analysis from Section 4.1, but this time with system-produced substitutes. Where necessary, we replaced substitute frequencies obtained for human-produced lexical substitutes with the model's substitute probabilities. The results are shown in the right half of Table 3. Comparing these correlations with those obtained with human-produced substitutes, we observe that using system-produced substitutes generally results in decreased correlations (0.076 on average). The difference is most pronounced for WordNet sense inventories: 0.071 and 0.079 for the best-performing measures on WN-Single and WN-Multi, respectively. Another observation is that the decrease in correlation is smaller for OntoNotes than for WordNet. In particular, the difference is negligible for the similarity measures that on OntoNotes perform the best, ParaDice and ParaCosineBin.

Taken together, the results indicate that there is indeed a performance gap between the human- and system-produced lexical substitutes, but its magnitude depends on the sense inventory used.

5. Word sense induction with lexical substitutes

The above experiments have shown that word meaning represented with human-produced lexical substitutes strongly correlates with sense-based meaning, but that the current performance gap is relatively large. This, however, does not entail that the gap will translate to downstream NLP tasks: it is conceivable that some NLP tasks will be robust to these differences, shrinking or even eliminating the performance gap. While here one could consider any of the numerous NLP tasks that build on word meaning representations, WSI seems the most natural choice given that it is a fundamental semantic task and also one that provides a direct link between substitute- and sense-based meaning representation. We therefore set to explore to what extent the decrease in correlation translates to performance on the WSI task (RQ3).

For this experiment, we consider (1) a number of simple WSI algorithms based on off-the-shelf clustering algorithms into which we incorporate our paraset-based similarity measures and (2) a state-of-the-art WSI model introduced in [12]. We next describe the WSI models, followed by experimental setup and results.

5.1. Simple WSI models

To avoid introducing algorithmic biases that could obscure our findings, we chose to rely on two standard and widely used clustering algorithms that operate on instance pair (dis-)similarities: hierarchical agglomerative clustering (HAC) and affinity propagation (AP). Both algorithms have been used for WSI and yielded satisfactory results [12,30,35]. We feed the algorithms with a (dis-)similarity matrix, computed by applying a particular paraset-based similarity measure across all instance pairs. Note that this design choice is in line with the similarity correlation analysis from Section 4.1, which used the same pairwise setup.

We used the readily-available implementations of *scikit-learn* [37] for both HAC and AP. We kept all hyperparameters at their respective default values, and only modified the number of clusters k for the HAC algorithm (which requires setting it upfront). For this value, following [12,35], we used the information about the actual number of senses: we set k to the average (AVG) and the maximum (MAX) number of word senses in the dataset (for each set of word sense annotations separately). We used ParaDice, ParaCosine, and ParaCosineBin as similarity measures.

5.2. State-of-the-art WSI model

The state-of-the-art WSI model introduced in [12] builds on the clustering-based approach from [35], which uses system-produced lexical substitutes. In this model, each word instance is associated with R representatives, each of which is composed of N lemmatized lexical substitutes sampled with replacement using LSDP with BERT (cf. Section 4.2). The sampled lexical substitutes are used to encode the representatives as one-hot vectors, which are then used to represent individual instances. Finally, all instances of a particular target word are encoded in this way and the resulting set of vectors is tf-idf weighted and clustered using HAC with cosine distance and average linkage. The process is repeated for every target word separately.

The described method results in a predefined number of hard clusters (i.e. each occurrence of a word belongs only to a single sense from the sense inventory). Soft clusters are obtained heuristically: the probability of an instance belonging to a cluster is obtained by computing the proportion of instance's representatives assigned to that particular cluster. Additionally, to sidestep the problem of using a fixed number of sense clusters (as in [35]), the method starts with a relatively high number of clusters and then merges them heuristically. More specifically, after clustering into 10 clusters (i.e. senses), the most probable sense for each of the instances is computed, and "weak senses" are identified as those that occurred fewer than two times. The weak senses are then merged with the closest non-weak sense based on the cosine distance between their centroids. If there were any merges, the soft cluster assignment is repeated. We do not consider soft clustering in our experiments and leave it for future work.

5.3. Model evaluation

The de-facto standard for WSI evaluation is the setup introduced in SemEval-2010, which we also adopt here. Although the original setup comprised two types of evaluation – supervised and unsupervised – following recent work [12] we focus on the unsupervised measures, V-measure [38] and paired F-measure [39]. As these metrics are suitable only for single-sense annotations, we evaluated our models only on the OntoNotes and WN-Single datasets.

We first describe the evaluation metrics. Let w be a target word with N instances (contexts). The instances are labelled with a set of sense labels $C = \{c_j \mid j = 1, 2, \ldots, n\}$ and are clustered into a number of clusters $K = \{k_j \mid j = 1, 2, \ldots, m\}$. Let $A = \{a_{ij}\}$ be the contingency matrix representing the clustering solution, such that a_{ij} denotes the number of instances labelled as c_i that belong to cluster k_j.

V-measure. This measure is a trade-off between two clustering properties: homogeneity and completeness. A clustering is *homogeneous* if all of its clusters contain only the instances of the same sense, while it is *complete* if all instances of the same sense belong to the same cluster. Homogeneity and completeness are measured via conditional entropy of the sense distribution given a clustering and via conditional entropy of the cluster distribution given a sense, respectively. Formally, homogeneity h is defined as:

$$h = \begin{cases} 1 & \text{if } H(C, K) = 0, \\ 1 - \dfrac{H(C \mid K)}{H(C)} & \text{otherwise.} \end{cases} \tag{1}$$

where $H(C \mid K)$ and $H(C)$ are defined as:

$$\begin{aligned} H(C \mid K) &= -\sum_{k=1}^{|K|} \sum_{c=1}^{|C|} \frac{a_{ck}}{N} \log \frac{a_{ck}}{\sum_{c=1}^{|C|} a_{ck}}, \\ H(C) &= -\sum_{c=1}^{|C|} \frac{\sum_{k=1}^{|K|} a_{ck}}{n} \log \frac{\sum_{k=1}^{|K|} a_{ck}}{n} \end{aligned} \tag{2}$$

Symmetrically, completeness c is defined as:

$$c = \begin{cases} 1 & \text{if } H(K, C) = 0, \\ 1 - \dfrac{H(K \mid C)}{H(K)} & \text{otherwise.} \end{cases} \tag{3}$$

where $H(K \mid C)$ and $H(K)$ are defined as:

$$H(K \mid C) = -\sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \frac{a_{ck}}{N} \log \frac{a_{ck}}{\sum_{k=1}^{|K|} a_{ck}},$$

$$H(K) = -\sum_{k=1}^{|K|} \frac{\sum_{c=1}^{|C|} a_{ck}}{n} \log \frac{\sum_{c=1}^{|C|} a_{ck}}{n} \qquad (4)$$

Finally, V-measure is calculated as the harmonic mean of homogeneity h and completeness c.

Paired F-measure. This measure is a clustering counterpart to the regular F-measure [40]. First, we generate within-cluster instance pairs for every predicted and reference cluster. Let $F(K)$ be the set of instance pairs generated from the predicted clusters, and $F(C)$ analogously for the reference clusters. Then, paired F-measure F is defined as a harmonic mean of precision P and recall R:

$$P = \frac{|F(K) \cap F(C)|}{|F(K)|}, \quad R = \frac{|F(K) \cap F(C)|}{|F(C)|},$$

$$F = \frac{2 \cdot PR}{P + R} \qquad (5)$$

Even though both V-measure and paired F-measure are devised as a trade-off between competing clustering properties, they should not be considered in isolation as they favour different clusterings. More concretely, the V-measure favours solutions with many clusters, whereas the paired F-measures penalizes them. To account for this, most studies report a geometric mean of the V-measure and the F-measure. We adopt the same metric here. To account for model non-determinism, we run the evaluation 10 times and average the scores.

5.4. Results

Table 4 shows the performance of the WSI models as the geometric mean of V-measure and F-measure for the different paraset similarity measures. The main observation pertains to RQ3: using human-produced lexical substitutes (the left half of the Table) still offers performance improvement over using system-produced substitutes (the right half of the table). However, the performance gap is modest (up to 0.15), in line with the results from Section 4.2. The paraset similarity measures perform similarly across all setups, with the exception of ParaCosine when dealing with system-produced lexical substitutes. When it comes to sense annotations, we again observe that WordNet-based annotations perform better, which was also shown in the similarity correlation analysis.

Regarding the choice of the clustering algorithm, HAC emerged as the clear winner, at least when using default hyperparameters. Still, using AP does not yield much worse performance and may be a satisfactory option when information about the number of sense clusters is not available (which is a more realistic setup).

When comparing simple WSI models to the state-of-the-art WSI model, two observations are pertinent. First, when using human-produced lexical substitutes, the best-performing simple model for a sense inventory slightly overcomes the respective state-of-the-art model's performance. Second, when switching to the system-produced lexical substitutes, this ceases to be the case. However, the obtained performance is still in the same ballpark as that of the state-of-the-art model, which indicates that the state-of-the-art model may be overly complex. Admittedly, some of this complexity serves to obtain soft sense labels, something that our models cannot tackle.

Table 4. WSI performance (geometric mean of V-measure and F-measure) for different models and sense inventories with human- and system-produced substitutes.

	Human		System	
Model	ON	WN	ON	WN
State-of-the-art model [12]	–	–	**0.538**	**0.600**
HAC |ParaDice |AVG	**0.546**	0.565	0.537	0.591
HAC |ParaDice |MAX	0.544	0.561	0.526	0.554
HAC |ParaCosine |AVG	0.533	**0.587**	0.458	0.494
HAC |ParaCosine |MAX	0.538	0.585	0.474	0.439
HAC |ParaCosineBin |AVG	0.543	0.573	0.534	0.581
HAC |ParaCosineBin |MAX	0.538	0.558	0.526	0.549
AP |ParaDice	0.434	0.476	0.415	0.440
AP |ParaCosine	0.449	0.529	0.362	0.379
AP |ParaCosineBin	0.432	0.475	0.415	0.440

Note: AVG = average number of clusters, MAX = maximum number of clusters. Best performances in each setup are shown in bold.

In sum, this experiment has demonstrated that, while there is a difference between human- and system-produced substitutes when used for the WSI task, the performance gap is relatively small and probably not of practical significance. This means that, for the task of WSI, representing word meaning with system-produced substitutes yields comparable performance as when representing word meaning using human-produced substitutes, which in turn justifies the use of automated LS for WSI.

6. Conclusion

Recent work in natural language processing has seen increased use of lexical substitution (LS) for representing word meaning in context, but it is not obvious how this representation corresponds to the more established sense-based meaning representation. This paper presented an empirical study of these questions. We found that there is a substantial positive correlation between substitute-based similarity and senses, contributing to the validity of the use of LS for word meaning representation. We also found that this correlation is generally lower for system-produced substitutes, but that the performance gap depends on the sense inventory used. Interestingly, we found that this performance gap mostly diminishes when system-produced substitutes are used for the WSI task, even with simple WSI models, justifying the use of automated LS for WSI.

There are a number of directions for future work, such as extending the study to more datasets and languages as well as investigating soft clustering approaches for graded-sense WSI. The treatment of multiword expressions and the interaction between context-based and substitute-based representations also merit further investigation.

Notes

1. This is tantamount to using Pearson's correlation coefficient with sense similarity encoded as 0,1.
2. The study in [12] used the top 200 words as a substitute pool from which their approach sampled the substitutes (cf. Section 5.2). We decided to keep this number the same.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

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

References

[1] Geeraerts D. Theories of lexical semantics. Oxford, United Kingdom: Oxford University Press; 2010.

[2] Miller GA. WordNet: a lexical database for English. Commun ACM. 1995;38(11):39–41.

[3] McCrae JP, Rademaker A, Rudnicka E, et al. English WordNet 2020: improving and extending a WordNet for english using an open-source methodology. In: Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020); Marseille, France; 2020. p. 14–19.

[4] Emerson G. What are the goals of distributional semantics? 2020. In press.

[5] Harris ZS. Distributional structure. Word. 1954;10 (2–3):146–162.

[6] Goldberg Y. Neural network methods for natural language processing. Synth Lect Hum Lang Technol. 2017;10(1):1–309.

[7] Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. Preprint, 2018. arXiv:181004805.

[8] McCarthy D, Navigli R. The English lexical substitution task. Lang Resour Eval. 2009;43(2):139–159.

[9] Zhou W, Ge T, Xu K, et al. BERT-based lexical substitution. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Florence, Italy; 2019. p. 3368–3373.

[10] Arefyev N, Sheludko B, Podolskiy A, et al. A comparative study of lexical substitution approaches based on neural language models. Preprint, 2020. arXiv:200600031.

[11] Melamud O, McClosky D, Patwardhan S, et al. The role of context types and dimensionality in learning word embeddings. Preprint, 2016. arXiv:160100893.

[12] Amrami A, Goldberg Y. Towards better substitution-based word sense induction. Preprint, 2019. arXiv: 190512598.

[13] Navigli R. A quick tour of word sense disambiguation, induction and related approaches. In: International Conference on Current Trends in Theory and Practice of Computer Science. Springer; 2012. p. 115–129.

[14] Schütze H. Automatic word sense discrimination. Comput Linguist. 1998;24(1):97–123.

[15] Hovy E, Marcus M, Palmer M, et al. OntoNotes: the 90% solution. In: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers; New York City, USA; 2006. p. 57–60.

[16] Szarvas G, Biemann C, Gurevych I, et al. Supervised all-words lexical substitution using delexicalized features. In: Proceedings of HLT-NAACL; Atlanta, Georgia; 2013. p. 1131–1141.

[17] Hintz G, Biemann C. Delexicalized supervised german lexical substitution. In: Proceedings of the GermEval GSCL Workshop; Essen, Germany; 2015. p. 11–16.

[18] Melamud O, Levy O, Dagan I. A simple word embedding model for lexical substitution. In: Proceedings of the NAACL VSM-NLP Workshop; Denver, Colorado; 2015. p. 1–7.

[19] Roller S, Erk K. PIC a different word: a simple model for lexical substitution in context. In: Proceedings of NAACL; San Diego, VA; 2016. p. 1121–1126.

[20] Kremer G, Erk K, Padó S, et al. What substitutes tell us – analysis of an "all-words" lexical substitution corpus. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics; Gothenburg, Sweden; 2014. p. 540–549.

[21] Biemann C. Turk bootstrap word sense inventory 2.0: a large-scale resource for lexical substitution. In: LREC; Istanbul, Turkey; 2012. p. 4038–4042.

[22] Buljan M, Padó S, Šnajder J. Lexical substitution for evaluating compositional distributional models. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers); New Orleans, Louisiana, USA; 2018. p. 206–211.

[23] McCarthy D, Apidianaki M, Erk K. Word sense clustering and clusterability. Comput Linguist. 2016;42(2): 245–275.

[24] Korkontzelos I, Manandhar S. UoY: graphs of unambiguous vertices for word sense induction and disambiguation. In: Proceedings of SemEval; Uppsala, Sweden; 2010. p. 355–358.

[25] Hope D, Keller B. UoS: a graph-based system for graded word sense induction. In: Proceedings of SemEval; Atlanta, Georgia; 2013. p. 689–694.

[26] Navigli R, Crisafulli G. Inducing word senses to improve web search result clustering. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing; Cambridge, Massachusetts, USA; 2010. p. 116–126.

[27] Bartunov S, Kondrashkin D, Osokin A, et al. Breaking sticks and ambiguities with adaptive skip-gram. Preprint, 2015. p. 47–54. arXiv:150207257.

[28] Komninos A, Manandhar S. Structured generative models of continuous features for word sense induction. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers; Osaka, Japan; 2016. p. 3577–3587.

[29] Başkaya O, Sert E, Cirik V, et al. AI-KU: using substitute vectors and co-occurrence modeling for word sense induction and disambiguation. In: Proceedings of SemEval; Atlanta, Georgia; 2013. p. 300–306.

[30] Alagić D, Šnajder J, Padó S. Leveraging lexical substitutes for unsupervised word sense induction. In: Thirty-Second AAAI Conference on Artificial Intelligence; Louisiana, New Orleans, USA; 2018. p. 5004–5011

[31] Manandhar S, Klapaftis I, Dligach D, et al. SemEval-2010 task 14: word sense induction & disambiguation. In: Proceedings of the 5th International Workshop on Semantic Evaluation; Uppsala, Sweden; 2010. p. 63–68.

[32] Jurgens D, Klapaftis I. SemEval-2013 task 13: word sense induction for graded and non-graded senses. In: Second Joint Conference on Lexical and Computational Semantics (∗ SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013); Atlanta, Georgia, USA; 2013. p. 290–299.

[33] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33: 159–174.

[34] Kishida K. Property of average precision and its generalization: an examination of evaluation indicator for information retrieval experiments. National Institute of Informatics Tokyo, Japan; 2005.

[35] Amrami A, Goldberg Y. Word sense induction with neural BiLM and symmetric patterns. Preprint, 2018. arXiv:180808518.

[36] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems; Long Beach, California, USA; 2017. p. 5998–6008.

[37] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–2830.

[38] Rosenberg A, Hirschberg J. V-measure: a conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL); Prague, Czech Republic; 2007. p. 410–420.

[39] Artiles J, Amigó E, Gonzalo J. The role of named entities in web people search. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing; Singapore; 2009. p. 534–542.

[40] van Rijsbergen C. Information retrieval. Oxford, United Kingdom: Butterworth-Heinemann; 1979.