Large Scale Substitution-based Word Sense Induction

Matan Eyal[1] **Shoval Sadde**[1] **Hillel Taub-Tabib**[1] **Yoav Goldberg**[1,2]
[1] Allen Institute for AI, Israel
[2] Bar Ilan University, Ramat-Gan, Israel
matane,shovals,hillelt,yoavg@allenai.org

Abstract

We present a word-sense induction method based on pre-trained masked language models (MLMs), which can cheaply scale to large vocabularies and large corpora. The result is a corpus which is sense-tagged according to a corpus-derived sense inventory and where each sense is associated with indicative words. Evaluation on English Wikipedia that was sense-tagged using our method shows that both the induced senses, and the per-instance sense assignment, are of high quality even compared to WSD methods, such as Babelfy. Furthermore, by training a static word embeddings algorithm on the sense-tagged corpus, we obtain high-quality static senseful embeddings. These outperform existing senseful embeddings methods on the WiC dataset and on a new outlier detection dataset we developed. The data driven nature of the algorithm allows to induce corpora-specific senses, which may not appear in standard sense inventories, as we demonstrate using a case study on the scientific domain.

1 Introduction

Word forms are ambiguous, and derive meaning from the context in which they appear. For example, the form "bass" can refer to a musical instrument, a low-frequency sound, a type of voice, or a kind of fish. The correct reference is determined by the surrounding linguistic context. Traditionally, this kind of ambiguity was dealt via *word sense disambiguation* (WSD), a task that disambiguates word forms in context between symbolic sense-ids from a sense inventory such as WordNet (Miller, 1992) or, more recently, BabelNet (Navigli and Ponzetto, 2010). Such sense inventories rely heavily on manual curation, are labor intensive to produce, are not available in specialized domains and inherently unsuitable for words with emerging senses.[1] This can be remedied by *word sense induction* (WSI), a task where the input is a given word-type and a corpus, and the output is a derived sense inventory for that word. Then, sense disambiguation can be performed over the WSI-derived senses.

The introduction of large-scale pre-trained LMs and Masked LMs (MLM) seemingly made WSI/WSD tasks obsolete: instead of representing tokens with symbols that encode sense information, each token is associated with a contextualized vector embeddings that captures various aspects of its in-context semantics, including the word-sense. These contextualized vectors proved to be very effective as features for downstream NLP tasks. However, contextualized embeddings also have some major shortcomings: most notably for our case, they are expensive to store (*e.g.* BERT embeddings are 768 or 1024 floating point numbers for each token), and are hard to index and query at scale. Even if we do manage to store and query them, they are not interpretable, making it impossible for a user to query for a particular sense of a word without providing a full disambiguating context for that word. For example, consider a user wishing to query a dataset for sentences discussing *Oracle* in the mythology-prophet sense, rather than the tech company sense. It is not clear how to formulate such a query to an index of contextualized word vectors. However, it is trivial to do for an index that annotates each token with its derived sense-id (in terms of UI, after a user issues a query such as "Oracle", the system may show a prompt such as "did you mean Oracle related to IBM; Sun; Microsoft, or to Prophet; Temple; Queen", allowing to narrow the search in the right direction).

Amrami and Goldberg (2018, 2019) show how contextualized embeddings can be used for achieving state-of-the-art WSI results. The core idea of their WSI algorithm is based on the intuition, first proposed by Başkaya et al. (2013), that occurrences of a word that share a sense, also share in-context

[1] For example, in current WordNet version, *Corona* has 6 synsets, none of them relates to the novel *Coronavirus*.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4738 - 4752

bug

Representatives					Neighbours				
bug$_0$	**bug**$_1$	**bug**$_2$	**bug**$_3$	**bug**$_4$	**bug**$_0$	**bug**$_1$	**bug**$_2$	**bug**$_3$	**bug**$_4$
insect	problem	feature	bomb	virus	bugs$_0$	vulnerability$_2$	bugs$_1$	bugs$_3$	flu$_2$
fly	flaws	fix	device	infection	beetle$_0$	glitch	patches$_2$	dumpster	staph
beetle	hole	code	bite	crisis	spider$_0$	rootkit	bug$_1$	laptop$_1$	hangover
Bugs	patch	dog	screen	disease	snake$_1$	bugs$_1$	updates$_1$	footage$_1$	nosebleed
worm	mistake	software	tag	surprise	worm$_0$	virus$_2$	patch$_2$	cruiser$_3$	pain$_4$

Java

Representatives		Neighbours	
Java$_0$	**Java**$_1$	**Java**$_0$	**Java**$_1$
Jakarta	Eclipse	Timor$_0$	Python$_0$
Indonesia	Jo	Sumatra$_1$	JavaScript
Bali	Apache	Sulawesi	Pascal$_2$
Indies	software	Sumatra$_0$	SQL
Holland	Ruby	Kalimantan	library$_3$

chair

Representatives		Neighbours	
chair$_0$	**chair**$_1$	**chair**$_0$	**chair**$_1$
head	seat	Chair$_0$	stool$_0$
chairman	position	chairperson	podium$_2$
president	wheelchair	chairman$_0$	desk$_0$
presided	professor	president$_0$	professorship
lead	table	Chairman$_0$	throne$_1$

pound

Representatives			Neighbours		
pound$_0$	**pound**$_1$	**pound**$_2$	**pound**$_0$	**pound**$_1$	**pound**$_2$
lb	dollar	beat	lb$_0$	rupee	smash$_2$
foot	marks	punch	pounds$_0$	shilling	kick$_1$
weight	coin	pump	lbs$_0$	dollar$_1$	stomp
ton	Mark	crush	ton$_2$	franc	slash$_0$
kilograms	mile	attack	lbs$_1$	penny$_0$	throw$_4$

train

Representatives		Neighbours	
train$_0$	**train**$_1$	**train**$_0$	**train**$_1$
training	railway	recruit$_0$	bus$_0$
prepare	track	equip	tram$_1$
educate	rail	recruit$_1$	trains$_1$
practice	line	volunteer$_2$	carriage$_0$
qualified	railroad	retrain	coach$_3$

Figure 1: Examples of induced word-senses for various words. For each sense we list the top-5 representatives, as well as the 5 closest neighbours in the static embeddings space.

substitutes. An MLM is then used to derive top-k word substitutes for each word, and these *substitute-vectors* are clustered to derive word senses.

Our main contribution in this work is proposing a method that scales up Amrami and Goldberg (2018)'s work to *efficiently* annotate all tokens in a large corpus (e.g. Wikipedia) with automatically derived word-senses. This combines the high-accuracy of the MLM-based approach, with the symbolic representation provided by discrete sense annotations. The discrete annotations are interpretable (each sense is represented as a set of words), editable, indexable and searchable using standard IR techniques. We show two applications of the discrete annotations, the first one is sense-aware information retrieval (§7), and the second is high-quality senseful *static* word embeddings we can derive by training a static embeddings model on the large sense annotated corpus (§8).

We first show how the method proposed by Amrami and Goldberg (2018) can be adapted from deriving senses of individual lemmas to efficiently and cheaply annotating *all the corpus occurrences* of *all the words in a large vocabulary* (§3). Deriving word-sense clusters for all of English Wikipedia words that appear as single-token words in BERT-LARGE's (Devlin et al., 2019) vocabulary, and assigning a sense to each occurrence in the corpus, required 100 hours of cheap P100 GPUs (5 hours of wall-clock time on 20 single GPU machines) followed by roughly 4 hours on a single 96-cores CPU machines. The whole process requires less than 50GB of disk space, and costs less than 150$ on Google Cloud platform.

After describing the clustering algorithm (§4), we evaluate the quality of our system and of the automatic sense tagging using SemEval datasets and a new manually annotated dataset we created (§5). We show that with the produced annotated corpora it is easy to serve sense-aware information retrieval applications (§7). Another immediate application is feeding the sense-annotated corpora to a static embedding algorithm such as word2vec (Mikolov et al., 2013), for deriving *sense-aware static embeddings* (§8). This results in state-of-the-art sense-aware embeddings, which we evaluate both on an existing WiC benchmark (Pilehvar and Camacho-Collados, 2019) and on a new challenging benchmark which we create (§9).

In contrast to WSD which relies on curated sense inventories, our method is data-driven, therefore resulting senses are corpus dependent. The method can be applied to any domain for which a BERT-like model is available, as we demonstrate by applying it to the PubMed Abstracts of scientific papers, using SCIBERT (Beltagy et al., 2019). The resulting senses cover scientific terms which are not typically found in standard sense inventories (§6).

Figure 1 shows examples of induced senses for selected words from the English Wikipedia corpus. For each sense we list 5 community-based representatives (§3), as well as the 5 closest neighbours in the sense-aware embedding space (§8). Additional examples are available in Appendix A. Code and resources are available in github.com/allenai/WSIatScale.

2 Related Work

Word Sense Induction and Disambiguation Previous challenges like Jurgens and Klapaftis (2013) focused on word sense induction for small sized datasets. To the best of our knowledge we are the first to perform large-scale *all-words* WSI. The closest work to our method is the substitution-based method proposed in Amrami and Goldberg (2018, 2019) which is the starting point to our paper. In that paper, the authors suggested a WSI algorithm designed for a small dataset (SemEval 2010, 2013) with a predefined set of ambiguous target words (See (§3) for more details on the algorithm). In our work, we change Amrami and Goldberg (2019) such that we can efficiently run sense induction on all the words in very large corpora.

An alternative approach for sense tagging is based on Word Sense Disambiguation (WSD). The two main WSD methods are Supervised WSD and Knowledge-based WSD. Supervised WSD suffers from the difficulty of obtaining an adequate amount of annotated data. Indeed, even SemCor, the largest manually annotated tagged corpus, consists of only 226,036 annotated tokens. Among different supervisied WSD methods, Zhong and Ng (2010) suggested a SVM based approach and Melamud et al. (2016); Yuan et al. (2016) suggested LSTMs paired with nearest neighbours classification. Knowledge-base WSD (Moro et al., 2014; Pasini and Navigli, 2017), on the other hand, avoids the reliance on large annotated word-to-sense corpus and instead maps words to senses from a closed sense inventory (*e.g.* WordNet (Miller, 1992), BabelNet (Navigli and Ponzetto, 2010)). As such, the quality of knowledge-based WSD heavily depends on the availability, quality and coverage of the associated annotated resources.

Sense Embeddings In §8 we exploit the sense-induced corpus to train sense embeddings. Reisinger and Mooney (2010) were the first to suggest creating multiple representations for ambiguous words. Numerous recent papers (Chen et al., 2014; Rothe and Schütze, 2015; Iacobacci et al., 2015; Pilehvar and Collier, 2016; Mancini et al., 2017; Iacobacci and Navigli, 2019) aim to produce similar embeddings, all of which use either WordNet or BabelNet as semantic network. Our method is similar to Iacobacci et al. (2015), with the difference being that they rely on semantic networks (via Babelfy (Moro et al., 2014)). In contrast and similarly to us, Pelevina et al. (2016) does not rely on lexical resources such as WordNet. The authors proposed splitting pretrained embeddings (such as word2vec) to a number of prototype sense-embeddings. Yet in our work, we directly learn the multi-prototype sense-embeddings which is only possible due to the large-scale corpus annotation. When comparing both methods in §9.1 we infer it is better to directly learn multi-prototype sense-embeddings.

3 Large Scale Sense Induction

3.1 Definition

We define large-scale sense induction as deriving sense clusters for all words in a large vocabulary and assigning a sense cluster to each corpus occurrence of these words.[2]

3.2 Algorithm

Contextualized BERT vectors contain sense information, and clustering the contextualized vectors results in sense clusters. However, storing a 1024 dimensional vector of 32bit floats for each relevant token in the English Wikipedia corpus requires over 8TB of disk-space, making the approach cumbersome and not-scalable. However, as shown by Amrami and Goldberg (2019), MLM based word-substitutes also contain the relevant semantic information, and are much cheaper to store: each word-id in BERTLARGE's vocabulary can be represented by 2 bytes, and storing the top-5 substitutes for each corpus position requires less than 20GB of storage space.[3]

[2] In *BERT-large-cased-whole-word-masking* this corresponds to 16k vocabulary items, that match to 1.59B full words in English Wikipedia, or 92% of all word occurrences. Analyzing the remaining words, only 0.01% appear in Wikipedia more than 100 times. We derive word senses to a substantial chunk of the vocabulary, which also corresponds to the most ambiguous words as less frequent words are substantially less polysemous (Hernández-Fernández et al., 2016; Fenk-Oczlon et al., 2010; Zipf, 1945).

[3] The size can be reduced further using adaptive encoding techniques that assign fewer bits to frequent words. We did not implement this in this work.

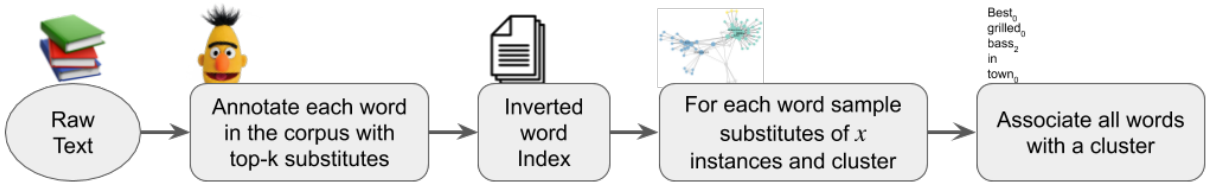

Figure 2: *Scalable WSI* flow. Given raw text, we annotate each word with its top-k substitutes, create inverted word index, find best clusters for each distinct lemma and associate all corpus words with a matching cluster.

In order to perform WSI at scale, we keep the main intuition from Amrami and Goldberg (2019), namely to cluster sparse vectors of lemmas of the top-k MLM-derived word substitutions. This results in vast storage saving, and also in a more interpretable representations. However, for scalability, we iterate over the corpus sentences and collect the top-k substitutes for all words in the sentence at once based on a single BERT call for that sentence. This precludes us from using the dynamic-patterns component of their method, which requires separately running BERT for each word in each sentence. However, as we show in Section §5.1 we still obtain sufficiently high WSI results.

The steps for performing Scalable WSI are summarized in Fig. 2. We elaborate on each step below, using English Wikipedia as a running example.[4]

Annotation: We run *BERT-large-cased-whole-word-masking* on English Wikipedia, inferring substitutes for all corpus positions. For positions that correspond to single-token words,[5] we consider the predicted words, filter stop-words, lemmatize the remaining words (Honnibal et al., 2020), and store the top-5 most probable lemmas to disk. This step takes 5 hours on 20 cloud-based GPU machines (total of 100 GPU hours), resulting in 1.63B tokens with their corresponding top-5 lemmas.

Inverted Word Index: We create an inverted index mapping from each single-token word to its corpus occurrences (and their corresponding top-5 lemmas). This takes 5 minutes on a 96 cores CPU machine, and 10GB of disk.

Sense Induction: For each of 16,081 lemmas corresponding to single-token words, we retrieve random 1000 instances,[6] and induce senses using the community-based algorithm described in §4. This process requires 30 minutes on the 96-core CPU machine, and uses 100MB of disk space. The average number of senses per lemma is 3.13. Each sense is associated with up to 100 representative words, which represent the highest-degree words in the sense's community. Table 1 shows the 5 senses found for the word *bass* with their top-5 representative words. See additional examples in Fig. 1 and Appendix A.

bass$_0$	**bass$_1$**	**bass$_2$**	**bass$_3$**	**bass$_4$**
bassist	double	fish	tenor	trap
guitar	second	bottom	baritone	swing
lead	tail	perch	voice	heavy
drum	steel	shark	soprano	dub
rhythm	electric	add	singer	dance

Table 1: Top 5 representatives of the sense-specific communities of word *bass*. The communities roughly match to bass as a musical instrument, register, fish species, voice and in the context of Drum&Bass

Tagging: Each of the remaining word-occurrences is associated with a sense cluster by computing the Jaccard similarity between the occurrences' top-5 lemmas and the cluster representatives, and choosing the cluster that maximizes this score. For example, an occurrence of the word *bass* with lemmas *tenor, baritone, lead, opera, soprano* will be associated with bass$_3$. This takes 100 minutes on 96-core machine, and 25GB of storage.

4 Sense Clustering Algorithm

We replace the hierarchical clustering algorithm used by Amrami and Goldberg (2018, 2019) with a community-detection, graph-based clustering algorithm. One major benefit of the community detection algorithms is that they naturally produces a dynamic number of clusters, and provide a list of interpretable discrete representative lemmas for each cluster. We additionally found this method to be more stable.

Graph-based clustering for word-sense induction typically constructs a graph from word occurrences

[4]The Wikipedia corpus is based on a dump from August 2020, with text extracted using WikiExtractor (Attardi, 2015).

[5]We exclude single-character tokens, stopwords and punctuation.

[6]The clustering algorithm scales super-linearly with the number of instances. To reduce computation cost for tokens that appear more than 1000 times in the dataset, we sample $\min(\text{numOccur}, 1000)$ instances for each token word, and cluster given the subset of instances. We then associate each of the remaining instances to one of the clusters as explained in the final step of the algorithm.

or collocations, where the goal is to identify sense-specific sub-graphs within the graph that best induce different senses (Klapaftis and Manandhar, 2008, 2010). We instead construct the graph based on word substitutes. Following Jurgens (2011), we pose identifying sense-specific clusters as a *community detection problem*, where a community is defined as a group of connected nodes that are more connected to each other than to the rest of the graph.

Graph construction For each word w in the vocabulary, we construct a graph $G_w = (V_w, E_w)$ where each vertex $v \in V_w$ is a substitute-word predicted by the MLM for w, and an edge $(u, v) \in E_w$ connects substitutes that are predicted for the same instance. The edge is weighted by the number of instances in which both u and v were predicted. More formally, let $X = \{x_w^i\}_{i=1}^n$ bet the set of all top-k substitutes for n instances of word w, and $x_w^i = \{{w'}^j_{x_w^i}\}_{j=1}^k$ represents the k top substitutes for the ith instance of word w. The graph G_w is defined as follows:

$$V_w = \{u : \exists i \; u \in x_w^i\}$$
$$E_w = \{(u, v) : \exists i \; u \in x_w^i \wedge v \in x_w^i\}$$
$$W(u, v) = |\{i : (u, v) \in x_w^i\}|$$

Community detection A community in a sub-graph corresponds to a set of tokens that tend to co-occur in top-k substitutes of many instances, and not co-occur with top-k substitutes of other instances. This corresponds well to senses and we take community's nodes as sense's representatives.

We identify communities using the fast *"Louvain"* method (Blondel et al., 2008). Briefly, Louvain searches for an assignment of nodes to clusters such that the *modularity score* Q—which measures the density of edges inside communities compared to edges between communities—is maximized:

$$Q = \frac{1}{2m} \sum_{u \; v} \left[W(u, v) - \frac{k_u k_v}{2m} \right] \delta(c_u, c_v)$$

m is the sum of all edge weights in the graph, $k_u = \sum_v W(u, v)$ is the sum of the weights of the edges attached to node u, c_u is the community to which u is assigned, and δ is Kronecker delta function. This objective is optimized using an iterative heuristic process. For details, see Blondel et al. (2008).

5 Intrinsic Evaluation of Clustering Algorithm

We start by *intrinsically evaluating* the WSI clustering method on: (a) SemEval 2010 and SemEval 2013; and (b) a new test set we develop for large-scale WSI. In section 9, we additionally *extrinsically evaluate* the accuracy of static embeddings derived from a sense-induced Wikipedia dataset.

When collecting word-substitutes, we lemmatize the top-k list, join equivalent lemmas, remove stopwords and the target word from the list, and keep the top-5 remaining lemmas.

5.1 SemEval Evaluation

We evaluate the community-based WSI algorithm on two WSI datasets: SemEval 2010 Task 14 (Manandhar et al., 2010) and SemEval 2013 Task 13 (Jurgens and Klapaftis, 2013). Table 2 compares our method to Amrami and Goldberg (2018, 2019) and AutoSense (Amplayo et al., 2019), which is the second-best available WSI method. Bert-noDP/DP are taken from Amrami and Goldberg (2019). Bert-DP uses "dynamic patterns" which precludes wide-scale application. We follow previous work (Manandhar et al., 2010; Komninos and Manandhar, 2016; Amrami and Goldberg, 2019) and evaluate SemEval 2010 using F-Score and V-Measure and SemEval 2013 using Fuzzy Normalized Mutual Information (FNMI) and Fuzzy B-Cubed (FBC) as well as their geometric mean (AVG). Our method performs best on SemEval 2010 and comparable to state-of-the-art results on SemEval 2013. The algorithm performs on-par with the Bert-noDP method, and does not fall far behind the Bert-DP method. We now turn to assess the end-to-end induction and tagging over Wikipedia.

5.2 Large Scale Manual Evaluation

We evaluate our method on large corpora by randomly sampling 2000 instances from the sense-induced Wikipedia, focusing on frequent words with many senses. We manually annotate the samples' senses without access to the automatically induced senses, and then compare our annotations to the system's sense assignments. We publicly release our manual sense annotations.

Sampling and Manual Annotation We used a list of 20 ambiguous words from *CoarseWSD-20* (Loureiro et al., 2021). The full list and per-word results can be found in Appendix C. For each word we sampled 100 passages from English Wikipedia

Model	F-S	V-M	AVG
AutoSense	61.7	9.8	24.59
Bert-noDP	70.9 (0.4)	37.8 (1.5)	51.7 (1.2)
Ours	**70.95 (0.63)**	**40.79 (0.19)**	**53.79 (0.31)**
Bert-DP	71.3 (0.1)	40.4 (1.8)	53.6 (1.2)

Model	FNMI	FBC	AVG
AutoSense	7.96	61.7	22.16
Bert-noDP	19.3 (0.7)	**63.6 (0.2)**	**35.1 (0.6)**
Ours	**19.42 (0.39)**	61.98 (0.12)	34.69 (0.33)
Bert-DP	21.4 (0.5)	64.0 (0.5)	37.0 (0.5)

Table 2: Evaluation on the SemEval 2010 (top) and SemEval 2013 (bottom) datasets. We report mean (STD) scores over 10 runs.

with the target word, including inflected forms (case insensitive). Unlike *CoarseWSD-20*, we sampled examples without any respect to a predefined set of senses. For example, the only two senses that appear in *CoarseWSD-20* for the target word *arm* are *arm (anatomy)*, and *arm (computing)*, leaving out instances matching senses reflecting *weapons, subdivisions, mechanical arms* etc.

With the notion that word sense induction systems should be robust to different annotations schemes, we gave two fluent English speakers 100 sentences for each of the 20 ambiguous words from *CoarseWSD-20*. Annotators were not given a sense inventory. Each annotator was asked to label each instance with the matching sense *according to their judgment*. For example, for the target word *apple* in the sentence *"The iPhone was announced by Apple CEO."*, annotators can label the target sense with *Apple Inc.*, *Apple The Company* etc. Annotation Guidelines are available in Appendix B.

On average annotators labeled 6.65 senses per word (5.85 and 7.45 average clusters per word for the two annotators). This is more than the 2.65 average senses according to *CoarseWSD-20* and less than WordNet's 9.85.

Results We report our system's performance alongside two additional methods: A strong baseline of the most frequent sense (MFS), and Babelfy (Moro et al., 2014)—the sense disambiguation system used in BabelNet (Tested using Babelfy live version April 2021). Differently from the latter, our system does not disambiguates but induces senses, therefore, clusters are not labeled with a sense tag from a sense inventory. Instead, we represent senses to annotators using a list of common substitute words and a few examples. Thus, after annotating the Wikipedia passages, we additionally asked annotators to name the system's clusters with the same naming convention as in their annotations.

	MFS	Babelfy	Ours
Ann #1	49.55	41.5	**89.05**
Ann #2	49.9	41.95	**85.95**
average	49.72	41.72	**87.50**

Table 3: Classification F1 scores for MFS, Babelfy and our proposed system by annotator on our manually annotated dataset.

Given a similar naming convention between systems and annotators, we report F1 scores of systems' tagging accuracy with respect to the manual annotations. We report F1 averaged over words in Table 3. Our system outperforms both baselines, despite Babelfy having access to a list of predefined word senses. A full by-word table and comprehensive results analysis are in Appendix C.

While a 1-to-1 mapping between system clusters and manual senses is optimal, our system sometimes splits senses into smaller clusters, thus annotators will name two system clusters with the same label. Therefore it is also important to report the number of clusters produced by the system comparing to the number of senses after the annotators merged similar clusters. Our system produced 7.25 clusters with 2.25 clusters on average merged by the annotators.[7] Additionally, in rare cases our system encapsulates a few senses in a single cluster: this happened 3 and 5 times for both annotators across all the dataset.

6 Application to Scientific Corpora

A benefit of a WSI approach compared to WSD methods is that it does not rely on a pre-specified sense inventory, and can be applied to any corpus for which a BERT-like model is available. Thus, in addition to the Wikipedia dataset that has been presented throughout the paper, we also automatically induce senses over a corpus of 31 million PubMed Abstracts,[8] using SciBERT (Beltagy et al., 2019). As this dataset is larger than the Wikipedia dump, the process required roughly 145 GPU hours and resulting in 14,225 sense-annotated lemmas, with an average number of 2.89 senses per lemma.

This dataset highlights the data-driven advantages of sense-induction: the algorithm recovers many senses that are science specific and are not represented in the Wikipedia corpora. While performing a wide-scale evaluation of the scientific WSI is beyond our scope in this work, we do show

[7] This is partially due to using clusters from two casing (*e.g. bank* and *Bank*), some of the merges share sense meaning but of different casing.

[8] www.nlm.nih.gov/databases/download/pubmed_medline

a few examples to qualitatively demonstrate the kinds of induced senses we get for scientific texts.

For each of the words *mosaic, race* and *swine* we show the induced clusters and the top-5 cluster representatives for each cluster.

mosaic$_0$	**mosaic**$_1$	**mosaic**$_2$	**mosaic**$_3$
virus	partial	mixture	mixed
dwarf	chimeric	landscape	genetic
mild	congenital	combination	spatial
cmv	heterozygous	pattern	functional
stripe	mutant	matrix	cellular

While senses mosaic$_0$ (the common mosaic virus of plants) and mosaic$_2$ ("something resembling a mosaic", "mosaic of..") are represented in Wikipedia, senses mosaic$_1$ (the mosaic genetic disorder) and mosaic$_3$ (mosaic is a quality, e.g., "mosaic border", "mosaic pattern") are specific to the scientific corpora (The Wikipedia corpora, on the other hand, includes a sense of mosaic as a decorative art-form, which is not represented in Pubmed).

race$_0$	**race**$_1$	**race**$_2$	**race**$_3$
racial	exercise	class	pcr
ethnicity	run	group	clone
black	training	state	sequence
rac	competition	population	rt
gender	sport	genotype	ra

Senses race$_0$ (ethnic group), race$_1$ (competition) and race$_2$ (population/civilization) are shared with wikipedia, while the sense race$_3$ ("Rapid amplification of cDNA ends", a technique for obtaining the sequence length of an RNA transcript using reverse transcription (RT) and PCR) is Pubmed-specific.

swine$_0$	**swine**$_1$	**swine**$_2$
pig	seasonal	patient
porcine	avian	infant
animal	influenza	group
livestock	pandemic	case
goat	bird	myocardium

Here swine$_1$ captures the Swine Influenza pandemic, while swine$_2$ refers to swine as experimental Pigs.

7 Sense-aware Information Retrieval

An immediate application of a high quality sense-tagged corpus is sense-aware retrieval. We incorporate the sense information in the SPIKE extractive search system (Shlain et al., 2020)[9] for Wikipedia and Pubmed datasets. When entering a search term, suffixing it with @ triggers sense selection allowing to narrow the search for the specific sense. Consider a scientist looking for PubMed occurrences of the word *"swine"* in its influenza meaning. As shown in Figure 3, this can be easily done by writing "swine@" and choosing the second item in the resulting popup window. The outputs are sentences with the word *"swine"* in the matching sense. As far as we know, SPIKE is the first system with such WSI capabilities for IR. Similarly, Blloshmi et al. (2021) suggested to enhance IR with sense information, but differently from us, this is done by automatically tagging words with senses from a predefined inventory.

8 Sense-aware Static Embeddings

Learning static word embeddings of sense-ambiguous words is a long standing research goal (Reisinger and Mooney, 2010; Huang et al., 2012). There are numerous real-world tasks where context is not available, precluding the use of contextualized-embeddings. These include Outlier Detection (Camacho-Collados and Navigli, 2016; Blair et al., 2016), Term Set Expansion (Roark and Charniak, 2000) the Hypernymy task (Breit et al., 2021), etc. Additionally, static embeddings are substantially more efficient to use, can accommodate larger vocabulary sizes, and can accommodate efficient indexing and retrieval. Yet, despite their flexibility and success, common word embedding methods still represent ambiguous words as a single vector, and suffer from the inability to distinguish between different meanings of a word (Camacho-Collados and Pilehvar, 2018).

Using our sense-tagged corpus we suggest a simple and effective method for deriving sense-aware static embeddings: We run an off-the-shelf embedding algorithm,[10] on the corpus where single-token words are replaced with a concatenation of the word and its induced sense (*e.g. "I caught a bass."* becomes *"I caught@0 a bass@2."*). This makes the embedding algorithm learn embeddings for all senses of each word out-of-the-box.[11] An integral property of the embedding algorithm is that it represents both the sense-annotated tokens and the other vocabulary items in the same embedding space —

[9] spike.apps.allenai.org

[10] We use the CBOW variant of the word2vec algorithm (Mikolov et al., 2013) as implemented in Gensim (Řehůřek and Sojka, 2010). We derive 100-dimensional embeddings using the negative-sampling algorithm and a window size of 5.

[11] A similar approach was used by Iacobacci et al. (2015) over a corpus which was labeled with BabelNet and WordNet senses.

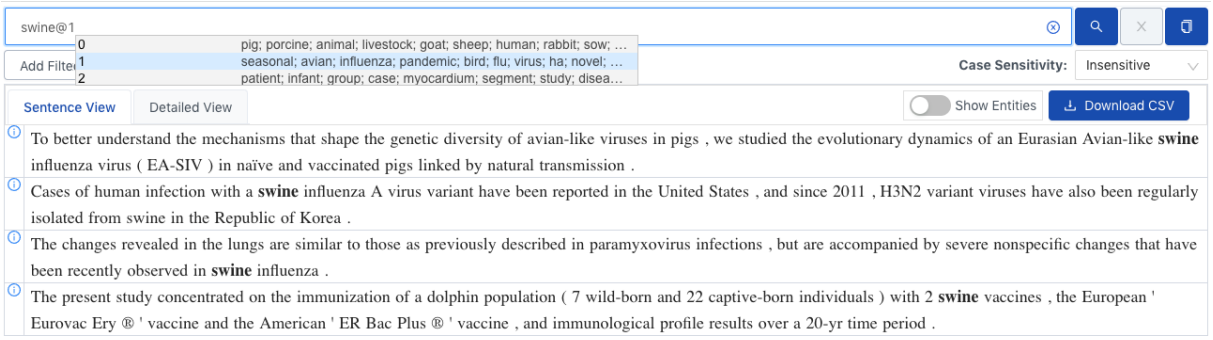

Figure 3: User interaction in SPIKE when looking for the word *"swine"* in its *"swine flu"* sense. (Unlike the animal/experimental pig senses)

this helps sense inferring about words that are represented in the MLM as multi-tokens words (Even though these correspond to less-frequent and often less ambiguous words (Hernández-Fernández et al., 2016; Fenk-Oczlon et al., 2010; Zipf, 1945)). For example, in the top-5 nearest neighbours for the different *bass* senses as shown below, *smallmouth* and *pumpkinseed*, multi-token words in BERTLARGE's vocabulary, are close neighbours the *bass* instances that correspond to the *fish* sense.

bass$_0$	**bass**$_1$	**bass**$_2$	**bass**$_3$	**bass**$_4$
guitar$_0$	tuba	crappie	baritone$_0$	synth
drums$_0$	trombone$_0$	smallmouth	tenor$_0$	drum$_1$
guitar$_3$	horn$_0$	pumpkinseed	alto$_0$	synths
keyboards$_0$	flute$_0$	sunfish	bassoon	breakbeats
keyboard$_0$	trumpet$_0$	perch$_0$	flute$_0$	trap$_4$

Note that some neighbours are sense annotated (single-token words that were tagged by our system), while others are not (multi-token words).

For English Wikipedia, we obtain a total vocabulary of 1.4M forms, 90,023 of which are sense-annotated. Compared to the community-based representative words, the top neighbours in the embedding space tend to capture members of the same semantic class rather than direct potential replacements.

9 Sense-aware Embeddings Evaluation

9.1 WiC Evaluation

Pilehvar and Camacho-Collados (2019) introduced the WiC dataset for the task of classifying word meaning in context. Each instance in WiC has a target word and two contexts in which it appears. The goal is to classify whether the word in the different contexts share the same meaning. *e.g.* given two contexts: *There's a lot of trash on the bed of the river* and *I keep a glass of water next to my bed when I sleep*, our method should return *False* as the sense of the target word *bed* is different.

Method	**Acc.**
JBT (Pelevina et al., 2016)	53.6
Sense-aware Embeddings (this work)	**58.3**
SW2V* (Mancini et al., 2017)	58.1
DeConf* (Pilehvar and Collier, 2016)	58.7
LessLex* (Colla et al., 2020)	**59.2**

Table 4: Accuracy scores on the WiC dataset. Systems marked with * make use of external lexical resources.

Word Embeddings	**OPP**	**Acc.**
GloVe	93.31	65
word2vec	93.31	68
DeConf	93.37	73
Ours (Skip-gram)	96.31	83.5
Ours (CBOW)	**96.68**	**86**

Table 5: OPP and Accuracy on the *25-7-1-8* dataset.

Our method is the following: Given the sense-aware embeddings, a target word w and two contexts, we calculate the context vector as the average of the context words. The matching sense vector is the closest out of all w embeddings. We then classify the contexts as corresponding to the same meaning if the cosine distance of the found sense embedding is more than threshold apart. We do not use the train set. The threshold is optimized over the development set and fixed to 0.68.

This task has a few tracks, we compare our embeddings systems to the best performing methods from the *Sense Representations* track. Of these, JBT (Pelevina et al., 2016), a lexical embedding method, is the only one that does not use an external lexical resource (induction). The results in Table 4 show accuracy on this task. We outperform the induction method, and are on-par with the lexicon-based methods, despite not using any external lexical resource.

9.2 Evaluation via Outlier Detection

Another setup for evaluating word embeddings is that of *outlier detection*: given a set of words, identify which one does not belong to the set (Blair

et al., 2016). Outlier detection instances are composed of in-group elements and a set of outliers from a related semantic space. In each evaluation round, one outlier is added to the in-group items, and the algorithm is tasked with finding the outlier. Existing outlier detection datasets either did not explicitly target sense-ambiguous words (*8-8-8* (Camacho-Collados and Navigli, 2016), *WikiSem500* (Blair et al., 2016)) or explicitly *removed* ambiguous words altogether (*25-8-8-sem* (Brink Andersen et al., 2020)).

Ambiguity-driven Outlier Detection. We construct a challenge set for outlier detection that specifically targets ambiguous cases. In order to account for sense ambiguity, we add a *distractor* to each of the in-group sets: the distractor is an item which has multiple senses, where the most salient sense does not belong to the group, while another sense does belong to the group. For example:

In-group: *zeus, hades, poseidon, aphrodite, ares, athena, artemis*
Outliers: *mercury, odysseus, jesus, sparta, delphi, rome, wrath, atlanta*
Distractor: *nike*

Here, a model which does not explicitly represent the greek-god sense of *nike* is likely to place it far away from the in-group instances, causing it to be mistakenly marked as the outlier.

The starting point for our dataset is *25-8-8-Sem* (Brink Andersen et al., 2020). This dataset contains 25 test groups, each with 8 in-group elements and 8 outliers, resulting in 200 unique test cases. The outliers are sorted in a decreasing degree of relatedness to the in-group elements. In our dataset we replace one of the in-group elements with an ambiguous distractor. For example, in the Greek-gods case above, we replaced the original 8$^{\text{th}}$ item (*"hera"*) with the ambiguous distractor *nike*. [12] The dataset consists of 25 groups of 7 non ambiguous group elements, 1 distractor and 8 outliers (*25-7-1-8*), similarly resulting 200 unique test cases.

Method Following Camacho-Collados and Navigli (2016), we rank each word likelihood of being the outlier by the average of all pair-wise semantic similarities of the words in $W \backslash \{w\}$. Therefore if w is an outlier, this score should be low. See Appendix D for additional details.

Metrics Camacho-Collados and Navigli (2016) proposed evaluating outlier detection using the accuracy (The fraction of correctly classified outliers among the total cases) and Outlier Position Percentage (OPP) metric. OPP indicates how close outliers are to being classified correctly:

$$OPP = \frac{\sum_{W \in D} \frac{OP(W)}{|W|-1}}{|D|} \times 100$$

where $OP(W)$ is the position of the outlier according to the algorithm.

Results In Table 5 we report performance of on the *25-7-1-8* set. Word2vec and GloVe accuracy scores are low while having high OPP scores. This is the expected behaviour for embeddings without sense awareness. These will position the distractor and the outlier furthest away from the group items while not designed to make the hard decision required for high Accuracy. Our sense-aware embeddings strongly outperform GloVe and word2vec which do not include senses. Our embeddings also outperform the word embeddings proposed in DeConf (Pilehvar and Collier, 2016), which are the best performing sense embeddings on WiC which are also publicly available.

10 Conclusion

We show that substitution-based word-sense induction algorithms based on word-substitutions derived from MLMs are easily scalable to large corpora and vocabulary sizes, allowing to efficiently obtain high-quality sense annotated corpora. We demonstrate the utility of such large-scale sense annotation, both in the context of a scientific search application, and for deriving high-quality sense-aware static word embeddings.

As a secondary contribution, we also develop a new variant of the Outlier Detection evaluation task, which explicitly targets ambiguous words.

11 Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement No. 802774 (iEXTRACT).

References

Reinald Kim Amplayo, Seung-won Hwang, and Min Song. 2019. Autosense model for word sense induction. In *The Thirty-Third AAAI Conference on Artifi-*

[12] We additionally changed terms that are debatably ambiguous and changed the *"African animals"* group to the more general *"animals"* as no distractors were found.

cial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 6212–6219. AAAI Press.

Asaf Amrami and Yoav Goldberg. 2018. Word sense induction with neural biLM and symmetric patterns. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4867, Brussels, Belgium. Association for Computational Linguistics.

Asaf Amrami and Yoav Goldberg. 2019. Towards better substitution-based word sense induction. *arXiv preprint arXiv:1905.12598*.

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Osman Başkaya, Enis Sert, Volkan Cirik, and Deniz Yuret. 2013. AI-KU: Using substitute vectors and co-occurrence modeling for word sense induction and disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 300–306, Atlanta, Georgia, USA. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Philip Blair, Yuval Merhav, and Joel Barry. 2016. Automated generation of multilingual clusters for the evaluation of distributed representations. *arXiv preprint arXiv:1611.01547*.

Rexhina Blloshmi, Tommaso Pasini, Niccolò Campolungo, Somnath Banerjee, Roberto Navigli, and Gabriella Pasi. 2021. Ir like a sir: Sense-enhanced information retrieval for multiple languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1030–1041.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Anna Breit, Artem Revenko, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. WiC-TSV: An evaluation benchmark for target sense verification of words in context. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1635–1645, Online. Association for Computational Linguistics.

Jesper Brink Andersen, Mikkel Bak Bertelsen, Mikkel Hørby Schou, Manuel R. Ciosici, and Ira Assent. 2020. One of these words is not like the other: a reproduction of outlier identification using non-contextual word representations. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 120–130, Online. Association for Computational Linguistics.

José Camacho-Collados and Roberto Navigli. 2016. Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 43–50, Berlin, Germany. Association for Computational Linguistics.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035, Doha, Qatar. Association for Computational Linguistics.

Davide Colla, Enrico Mensa, and Daniele P. Radicioni. 2020. LessLex: Linking multilingual embeddings to SenSe representations of LEXical items. *Computational Linguistics*, 46(2):289–333.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gertraud Fenk-Oczlon, August Fenk, and Pamela Faber. 2010. Frequency effects on the emergence of polysemy and homophony. *International Journal of Information Technologies and Knowledge*, 4(2):103–109.

Antoni Hernández-Fernández, Bernardino Casas, Ramon Ferrer-i Cancho, and Jaume Baixeries. 2016. Testing the robustness of laws of polysemy and brevity versus frequency. In *International Conference on Statistical Language and Speech Processing*, pages 19–29. Springer.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 873–882, Jeju Island, Korea. Association for Computational Linguistics.

Ignacio Iacobacci and Roberto Navigli. 2019. Lstmembed: Learning word and sense representations from a large semantically annotated corpus with long short-term memories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1685–1695.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105, Beijing, China. Association for Computational Linguistics.

David Jurgens. 2011. Word sense induction by community detection. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, pages 24–28, Portland, Oregon. Association for Computational Linguistics.

David Jurgens and Ioannis Klapaftis. 2013. 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)*, pages 290–299, Atlanta, Georgia, USA. Association for Computational Linguistics.

Ioannis Klapaftis and Suresh Manandhar. 2010. Word sense induction & disambiguation using hierarchical random graphs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 745–755, Cambridge, MA. Association for Computational Linguistics.

Ioannis P Klapaftis and Suresh Manandhar. 2008. Word sense induction using graphs of collocations. In *ECAI*, pages 298–302.

Alexandros Komninos and Suresh Manandhar. 2016. Structured generative models of continuous features for word sense induction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3577–3587.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, pages 1–55.

Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. 2010. SemEval-2010 task 14: Word sense induction &disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala, Sweden. Association for Computational Linguistics.

Massimiliano Mancini, Jose Camacho-Collados, Ignacio Iacobacci, and Roberto Navigli. 2017. Embedding words and senses together via joint knowledge-enhanced training. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 100–111, Vancouver, Canada. Association for Computational Linguistics.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Tommaso Pasini and Roberto Navigli. 2017. Train-O-Matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 78–88, Copenhagen, Denmark. Association for Computational Linguistics.

Maria Pelevina, Nikolay Arefiev, Chris Biemann, and Alexander Panchenko. 2016. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 174–183, Berlin, Germany. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Nigel Collier. 2016. De-conflated semantic representations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1680–1690, Austin, Texas. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Joseph Reisinger and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117, Los Angeles, California. Association for Computational Linguistics.

Brian Roark and Eugene Charniak. 2000. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. *arXiv preprint cs/0008026*.

Sascha Rothe and Hinrich Schütze. 2015. AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1793–1803, Beijing, China. Association for Computational Linguistics.

Micah Shlain, Hillel Taub-Tabib, Shoval Sadde, and Yoav Goldberg. 2020. Syntactic search by example. In *ACL*.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1374–1385, Osaka, Japan. The COLING 2016 Organizing Committee.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, Uppsala, Sweden. Association for Computational Linguistics.

George Kingsley Zipf. 1945. The meaning-frequency relationship of words. *The Journal of general psychology*, 33(2):251–256.

A Additional Examples

Due to limit of space we provide additional examples in the appendix. We start with the senses found for the word *face*:

Representatives

face$_0$	**face**$_1$	**face**$_2$	**face**$_3$
confront	head	look	side
meet	front	address	line
encounter	name	point	wall
suffer	cheek	serve	surface
experience	body	toward	slope

Neighbours

face$_0$	**face**$_1$	**face**$_2$	**face**$_3$
meet$_3$	hand$_0$	faced$_2$	slope$_0$
challenge$_3$	forehead$_0$	sit$_1$	rim$_0$
suffer$_0$	hands$_0$	hang$_1$	flank$_2$
confront$_0$	nose$_0$	facing$_2$	ridge$_4$
lose$_1$	eyes$_3$	rotate$_0$	slope$_1$

The face senses refer to meeting/confronting, the body part, turn/look and side, respectively.

Here we present two senses of the word *orange*, corresponding to the color and fruit:

Representatives		**Neighbours**	
orange$_0$	**orange**$_1$	**orange**$_0$	**orange**$_1$
yellow	apple	yellow$_0$	apple$_0$
red	lemon	purple$_0$	avocado
amber	lime	amber$_0$	almond
pink	fruit	blue$_0$	apple$_1$
olive	banana	orangish	apricot

Finally we present the senses for *Jordan*:

Representatives

Jordan$_0$	**Jordan**$_1$	**Jordan**$_2$	**Jordan**$_3$
Johnson	Jerusalem	David	River
Jones	Palestine	Jason	Zion
Jackson	Israel	Joel	Water
Murray	Yemen	Justin	City
Mason	Turkey	Jonathan	water

Neighbours

Jordan$_0$	**Jordan**$_1$	**Jordan**$_2$	**Jordan**$_3$
Jones$_1$	Kuwait$_1$	Jeremy$_1$	Huleh
Kramer$_1$	Lebanon$_0$	Aaron$_0$	Yarkon
Allen$_0$	Syria$_0$	Justin$_0$	Arabah
Mack$_0$	Iraq$_0$	Brandon$_0$	Khabur
Robinson$_0$	Sudan$_1$	Josh$_0$	Tyropoeon

Here the clusters correspond to Jordan the surname, the country, first name and the Jordan River, respectively.

B Annotation Guidelines for Manual Evaluation

The objective of this task is to annotate word-meanings of 20 ambiguous words in a total of 2000 different contexts.

What is word-meaning? Words have different meanings in different contexts, for example, in the sentence: *"there is a light that never goes out"*, the word *"light"* refers to any device serving as a source of illumination. While *"light"* in the sentence *"light as a feather"* refers to the comparatively little physical weight or density of an object.

Step 1:

In this dataset we examine 20 ambiguous words as targets. For each of these words we collected 100 sentences in which the target word appears. For every sentence in the 100 set per target word, you will be asked to write a short label expressing the meaning of the target word in that particular context.

For example, here are three sentences with the target word *"light"*, each with its possible annotation.

1. *"there is a light that never goes out"* → visible light.
2. *"light as a feather"* → light as in weight.
3. *"magnesium is a light metal"* → light as in weight.

Note that in this example the annotator found the second and third meanings of the word *"light"* to be the same and therefore labeled them with the same label.[13]

While some annotations are indeed intuitive, labeling word-meanings when the target word is part of a name can be challenging. Here are a few guidelines for such use case:

Whenever a target word appeared as part of a name (Person, Organization etc.), one of three heuristics should be used[14]:

1. If the target word is the surname of a person, the example should be tagged *surname*.[15]

2. If the entity (as a whole) refers to one of the word-meanings, it should be labeled as such. For example, *Quitobaquito Springs* label should refer to a natural source of water.

3. If the target word is part of a name different from the original word-meaning, it should be tagged as *Part of Name*. This includes song names, companies (*Cold Spring Ice*), restaurants etc. Possible exceptions for this case are when a specific named entity is significantly frequent.

Step 2: [16]

[13] For ease of use for future evaluators, at the end of this step, both annotators picked a single naming convention when two labels referred to the same sense. Names of labels that were used only by one annotator were not changed.

[14] Some of the dissimilarities between the annotations are with respect the tension between the second and third guidelines.

[15] As opposed to Babelfy, there was no attempt for entity linking, so all persons were tagged the same.

[16] This step is presented to annotators once step 1 is done

For each of the target words you labeled, you will now receive a short list of indirect word-meaning definitions. Indirect word-meanings are composed of:

(a) A list of 10 words that may appear instead of the target word in specific contexts

(b) A list of 5 sentences in which the target word has this specific word-meaning.

For example, this is a possible indirect word-meaning for the target word *"Apple"*, representing the fruit, as opposed to the tech company:

Alternatives: orange, olive, cherry, lime, banana, emerald, lemon, tomato, oak, arrow,

Sentences in which Apple appears in this word-meaning:

"He and his new bride planted apple trees to celebrate their marriage."

"While visiting, Luther offers Alice an apple."

"When she picks the apple up, it is revealed that Luther has stolen a swipe card and given it to Alice to help her escape."

You will be asked to label the indirect word-meanings with one of the labels you used in step 1. If no label matches the indirect word-meaning you are allowed to propose a new label or define it to be "Unknown". Additionally, if you find several indirect word-meanings too close in meaning, label them the same.

C Analysis of Manual Evaluation

In table 6 we report a by-word analysis of our manual evaluation results. For each word we detail F1 scores of the most frequent sense (MFS), Babelfy, and our proposed system. Similarly to Loureiro et al. (2021), we report the ratio of the first sense with respect to the rest (F2R) and normalized entropy[17] to reflect sense balance. All of which are reported per annotator.

Analysis Analysis of our system's error shows that for some words the system could not create a matching cluster for specific senses (to name a few examples, "yard" as a ship identifier and "impound/enclosure" sense for the word "pound"). It appears that a matching cluster was not created due to the low tally of these senses in the English Wikipedia, and indeed the two senses appeared only two and three times respectively in the 100 passages sample. Additionally, annotator 2 annotated in a more fine-grained manner that does not correspond to our system tendency to merge capitalized instances of the target word into a sense that corresponds to "part of named entity".

As described above, in rare cases our system merged two senses into a single cluster. For example, the same cluster of the word "trunk" contained occurrences which annotator 1 tagged either "human torso" or "tube-like organs" (like the pulmonary trunk). While such annotation was uncommon (3 out of 117 senses for annotator 1 and 5 out of 149 senses for annotator 2), it does affect our system's micro F1 score for the better. In case we do not allow such annotation our overall score drops from 87.52 to 86.65.

A comparison between Babelfy and our gold annotation shows a common mistake in its labeling where Babelfy attributes the vast majority of sentences to the same non-salient sense. For example, Babelfy attributes 77 out of 100 instances of *hood* to "An aggressive and violent young criminal" - a sense that was not found even once in the manual annotation. While in a number of cases Babelfy used finer-grained sysnset groups than in our annotations we took into account any senses that are a subset of our annotated senses. For examples, Babelfy's *"United States writer who lived in Europe; strongly influenced the development of modern literature (1885-1972)"* synset was attribute any instances from the senses *surname* that refer to the writer Ezra Pound.

D Outlier Detection Method

When using a single-prototype vector-space models, Camacho-Collados and Navigli (2016) proposed a procedure for detecting outliers based on semantic similarity using *compactness score*:

$$c(w) = \frac{1}{n^2 - n} \sum_{w_i \in W \setminus \{w\}} \sum_{\substack{w_j \in W \setminus \{w\} \\ w_i \neq w_j}} sim(w_i, w_j)$$

Where D is the entire dataset and W is defined as $\{w_1, w_2, \cdots, w_n, w_{n+1}\}$ where w.l.o.g. $\{w_1, w_2, \cdots, w_n\}$ are the group elements (including the distractor) and w_{n+1} is the outlier. We use the same procedure with an additional nuance, we expanded the procedure to receive more than a single vector representation per word such that it will fit multi-prototype embeddings (*e.g.* our embeddings and DeConf) and case sensitive embeddings

for all words

[17] Computed as $\frac{-\sum_{i=1}^{k} \frac{c_i}{n} \log \frac{c_i}{n}}{log(k)}$, where k is the number of annotated senses, each of size c_i and n is the size of annotated examples per word, in our case $n = 100$.

	Annotator #1					Annotator #2				
Word	MFS	Babelfy	Ours	F2R	Ent.	MFS	Babelfy	Ours	F2R	Ent.
Apple	48	69	**94**	0.92	0.71	47	66	**86**	0.89	0.05
Arm	34	31	**89**	0.52	0.87	34	33	**85**	0.52	0.83
Bank	48	61	**94**	0.92	0.78	46	61	**85**	0.85	0.69
Bass	61	6	**82**	1.56	0.64	65	17	**83**	1.86	0.62
Bow	31	14	**80**	0.45	0.80	32	16	**80**	0.47	0.83
Chair	66	29	**90**	1.94	0.66	67	31	**86**	2.03	0.63
Club	49	45	**80**	0.96	0.78	53	50	**77**	1.13	0.72
Crane	39	36	**86**	0.64	0.90	39	35	**83**	0.64	0.69
Deck	45	49	**72**	0.82	0.80	48	52	**71**	0.92	0.68
Digit	87	96	**99**	6.69	0.56	87	96	**98**	6.69	0.38
Hood	27	6	**82**	0.37	0.88	28	5	**82**	0.39	0.83
Java	63	32	**98**	1.70	0.67	63	31	**97**	1.70	0.69
Mole	37	32	**90**	0.59	0.81	39	32	**88**	0.64	0.73
Pitcher	95	**97**	**97**	19.00	0.20	95	**97**	**97**	19.00	0.20
Pound	46	58	**91**	0.85	0.75	46	58	**91**	0.85	0.72
Seal	30	48	**88**	0.43	0.91	27	40	**74**	0.37	0.80
Spring	57	0	**90**	1.33	0.63	56	0	**88**	1.27	0.64
Square	37	15	**88**	0.59	0.86	36	15	**85**	0.56	0.82
Trunk	33	46	**98**	0.49	0.90	33	46	**92**	0.49	0.86
Yard	58	60	**93**	1.38	0.63	57	58	**91**	1.33	0.59
Average	49.55	41.5	**89.05**	2.11	0.74	49.9	41.95	**85.95**	2.13	0.65

Table 6: Manually annotated set scores by annotator. The first three columns for each annotator reflect disambiguation and induction scores with respect to the most frequent sense, Babelfy and our proposed system. We also report F2R and normalized entropy (Ent).

(*e.g.* word2vec). When given as set of words (like $W \setminus \{w\}$ when calculating $c(w)$) we first find the relevant sense for each element before inferring the outlier. Camacho-Collados and Navigli (2016) suggested calculating $c(w)$ using the *pseudo inverted compactness score*.