Combining Lexical Substitutes in Neural Word Sense Induction

Nikolay Arefyev[1,2], **Boris Sheludko**[1,2], **and Alexander Panchenko**[3]

[1]Samsung R&D Institute Russia, Moscow, Russia
[2]Lomonosov Moscow State University, Moscow, Russia
[3]Skolkovo Institute of Science and Technology, Moscow, Russia
narefjev@cs.msu.ru, b.sheludko@samsung.com, a.panchenko@skoltech.ru

Abstract

Word Sense Induction (WSI) is the task of grouping of occurrences of an ambiguous word according to their meaning. In this work, we improve the approach to WSI proposed by Amrami and Goldberg (2018) based on clustering of lexical substitutes for an ambiguous word in a particular context obtained from neural language models. Namely, we propose methods for combining information from left and right context and similarity to the ambiguous word, which result in generating more accurate substitutes than the original approach. Our simple yet efficient improvement establishes a new state-of-the-art on WSI datasets for two languages. Besides, we show improvements to the original approach on a lexical substitution dataset.

1 Introduction

Ambiguity, including lexical ambiguity, is one of the fundamental properties of natural languages and is a central challenge for NLP and its applications. Lexical ambiguity is a common situation when a single word has several meanings which can be either closely related (*coffee* as a plant, as a drink, or as beans for preparing that drink) or entirely unrelated (*band* as a musical group or as a strip of material). Consider the word *book* in *book a flight* or *buy a book*. Depending on the expressed meaning, machine translation systems should translate this word differently, search engines should find different information, personal digital assistants should take different actions, etc.

Word sense induction (WSI) is the task of clustering of occurrences of an ambiguous word according to their meaning. For evaluation of WSI systems, text fragments containing ambiguous words are hand-labeled with senses from some sense inventory (a dictionary or a lexical ontology). WSI systems are given text fragments only and should cluster them into some a priori unknown number of clusters (unlike Word Sense Disambiguation systems, which are also given the sense inventory).

Words that can appear instead of an ambiguous word in a particular context, also known as lexical contextual substitutes, are very helpful for WSI because possible substitutes strongly depend on the expressed meaning of the ambiguous word. For instance, for the word *build* possible substitutes are *manufacture*, *make*, *assemble*, *ship*, *export* if it is used in the *manufacturing some goods* sense and *erect*, *rebuild*, *open* for the *constructing a building* sense. Baskaya et al. (2013) proposed generating substitutes using n-gram language models and had shown one of the best results at the SemEval-2013 WSI shared task for English (Jurgens and Klapaftis, 2013). Later Amrami and Goldberg (2018) proposed generating contextual substitutes with a bidirectional neural language model (biLM) ELMo (Peters et al., 2018). With several other improvements, they had achieved new state-of-the-art results on the same dataset. However, their method simply unites substitutes generated independently from probability distributions $P(w_i|w_{i-1}...w_1)$ and $P(w_i|w_{i+1}...w_T)$ estimated by the forward and the backward ELMo LM independently, each given only one-sided context. This results in noisy substitutes when either left or right context is short or non-informative.

The **main contribution** of this paper is an approach that combines the forward and the backward distributions into a single distribution and fuses the similarity to the ambiguous word into the combined distribution. This allows taking into

Proceedings of Recent Advances in Natural Language Processing, pages 62–70,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_008

account all information we have about a particular ambiguous word occurrence for better substitutes generation. We compare several methods for combining distributions. Substitutes retrieved from the combined distribution perform much better for WSI achieving the a state-of-the-art on the SemEval 2013 dataset for English as well two datasets for Russian.

2 Related Work

The first methods to word sense induction were proposed already in the late 90s (Pedersen and Bruce, 1997; Schütze, 1998; Lin, 1998) with several competitions being organized to systematically evaluate various methods, including SemEval 2007 task 2 (Agirre and Soroa, 2007), SemEval 2010 task 14 (Manandhar et al., 2010) and SemEval 2013 task 13 (Jurgens and Klapaftis, 2013) for the English language, and RUSSE 2018 (Panchenko et al., 2018) for the Russian language.[1] Navigli (2012) provides a survey of word sense induction and related approaches. Methods for word sense induction can be broadly classified into three groups: context clustering approaches, word (ego-network) clustering, and latent variable models. We discuss these approaches below. Also, note that methods for learning word sense embedding (Camacho-Collados and Pilehvar, 2018) can be used to induce vector representations of senses from text.

2.1 Context/Vector Clustering Methods

This methods from this group represent a word instance by a vector that characterizes its context, where the definition of context can vary greatly. These vectors are subsequently clustered.

Early approaches, such as (Pedersen and Bruce, 1997; Schütze, 1998; Reisinger and Mooney, 2010) used sparse vector representations. Later approaches dense vector representations were adopted, e.g. Arefyev et al. (2018) and Kutuzov (2018) used weighted word embeddings (Mikolov et al., 2013) pre-trained on a large corpus to represent context of an ambiguous target word. Anwar et al. (2019) used contextualized (Peters et al., 2018) and non-contextualized (Mikolov et al., 2013) word embeddings to cluster occurrences of ambiguous occurrences of verbs according to their semantic frames.

[1] https://russe.nlpub.org

The approach presented in this paper is also an instance of vector clustering methods. More specifically, it exploits contextual substitutes for the ambiguous word to differentiate between its senses. Baskaya et al. (2013) proposed using substitute vectors for WSI, and their system *AI-KU* was one of the best-performing systems at SemEval 2013. Alagić et al. (2018) proposed another approach which leverages lexical substitutes for unsupervised word sense induction. They perform clustering of contexts using the affinity propagation algorithm (Dueck and Frey, 2007). The similarity between instances is measured using three different measures based on cosine similarities between pre-trained word embeddings by Mikolov et al. (2013). One measure relies on an average of embeddings of context words. Another one relies on an average of embeddings of lexical substitutes (also combination of both measures is tested). Finally, Amrami and Goldberg (2018) proposed using neural language models and dynamic symmetric patterns establishing a new best result on this dataset. Their approach is described in details in Section 3 as a starting point for our method.

2.2 Word/Graph Clustering Methods

This group of methods cluster word ego-networks consisting of a single node (ego) together with the nodes they are connected to (alters) and all the edges among those alters. Nodes of an ego-network can be words semantically similar to the target word or context features relevant to the target. This line of work starts from the seminal work of (Widdows and Dorow, 2002) who used graph-based methods for unsupervised lexical acquisition. In this work, senses of the word were defined as connected components in a graph which excludes the ego. Véronis (2004), Biemann (2006), and Hope and Keller (2013) further developed this idea by performing clustering of nodes instead of the simple search for connected components. Pelevina et al. (2016) proposed to transform word embeddings to sense embeddings using graph clustering (Biemann, 2006). The obtained sense embeddings were used to solve the WSI task based on similarity computations between the context and the induced sense.

2.3 Latent Variable Methods

Methods from this group, define a generative process of the documents which include word senses as a latent variable and then perform estimation

of the model from unlabeled textual data. For instance, Lau et al. (2013) relies on the Hierarchical Dirichlet Process (HDP) (Teh et al., 2006). Latent topics discovered in the training instances, specific to every word, are interpreted as word senses. Since the HDP is generative, also new instances can be assigned a sense topic. Latent variable model of Bartunov et al. (2016) is a Bayesian extension of Skip-gram (Mikolov et al., 2013) that automatically learns the number of word senses; it relies on the stick-breaking process. Amplayo et al. (2019) propose another graphical model which tackles the sense granularity problem, setting new state-of-the-art results for the SemEval 2010/2013 WSI datasets.

3 Bayesian Fusion of Lexical Substitutes from Bidirectional Language Models

In this section, we describe the method of word sense induction proposed by Amrami and Goldberg (2018), which is based on lexical substitutes generated given left and right context separately and then united together. Then we propose several methods to build a combined distribution incorporating information from left and right context as well as the similarity to the target word for better substitutes generation. For qualitative comparison, Table 1 lists lexical substitutes generated by different methods for several randomly selected sentences from the TWSI dataset (Biemann, 2012). For readability, we select either the top 10 predictions from the combined distributions or the union of the top 5 predictions from the forward and the backward distributions. The actual number of substitutes may be smaller due to duplicates appearing after lemmatization of substitutes.

3.1 Baselines: No Fusion (Union of Substitutes)

We base our approach on the method by Amrami and Goldberg (2018) (named **original** hereafter), which has achieved state-of-the-art results on the SemEval-2013 dataset for English WSI. Suppose c is the target ambiguous word and l, r are its left and right contexts. First, the method employs pretrained forward and backward ELMo LMs (Peters et al., 2018) to estimate probabilities for each word w to be a substitute for c given only the left context $P_{fwd}(w|l)$ or only the right context $P_{bwd}(w|r)$. Second, from the top K most probable words of each distribution L substitutes are sampled. This is done S times resulting in S representatives of the original example consisting of $2L$ substitutes each. Then TF-IDF BoW vectors for all representatives of all examples of a particular ambiguous word are built. Finally, agglomerative clustering is performed on the obtained representations with a fixed number of clusters. To provide more information to the LMs the target word can be included in the context using the technique called dynamic patterns. For example, given the sentence *These apples are sold everywhere* instead of *'These _'* the forward LM receives *'These apples and _'* and instead of *'_ are sold everywhere'* the backward LM receives *'_ and apples are sold everywhere'*. The underscore denotes the position for which the language model predicts possible words.

Thus, lexical substitutes are obtained **independently** from the forward and the backward LM and then **united**. For soft clustering required by the SemEval-2013 dataset, the probability distribution over clusters for each example is estimated from the number of representatives of this example put in each cluster. For the RUSSE (the Russian WSI) datasets we further convert soft clustering into hard clustering by selecting the most probable cluster for each example.

The second baseline (named **base**) simplifies the **original** method by skipping sampling and using $S = 1$ representative consisting of the union of the top K predictions from each LM. While being simpler and deterministic, this modification also delivers better results on RUSSE. Additionally, we have found that baselines with dynamic patterns translated into Russian perform worse than their counterparts without patterns (**original-no-pat** and **base-no-pat**) on RUSSE. This is in line with the ablations study from Amrami and Goldberg (2018) who found that the patterns are useful for verbs and adjectives but almost useless for nouns which the RUSSE datasets consist of. Interestingly, our best models perform better without dynamic patterns on all datasets.

3.2 Fusion at the Level of LM Distributions

During preliminary experiments, we have found that uniting substitutes retrieved from the forward and the backward LM independently results in lots of substitutes not related to the target word sense. For instance, consider the first example in Table 1 where the ambiguous word is the last word of the sentence. The backward LM simply

base-no-pat	base	BComb-LMs	BComb-3
It offers courses at the Undergraduate and Post Graduate levels in various subjects.			
sept, industry, feb, university, discipline, nov, dec, language, field, oct	offer, **course**, teach, subject, style, **topic**, background, size, include, provide	profession, subject, industry, university, discipline, sector, guise, language, field, department	field, occupation, language, discipline, sector, guise, profession, subject, department, industry
Wakeboards with a three - stage rocker push more water in front of the wakeboard, making the ride slower but riders are able to jump higher off the water.			
slightly **trip** perfect become **journey** climb **trek** bit speed	faster landing climb bend rid harder speed walk bike	jump incline slope climb bend **trek**	dive incline climb **trek** slope **journey** crawl
The couple were married on the bride's family estate at Ballyhooly, Cork, Ireland; afterwards the couple set up home at Caddington Hall.			
tree bear **residence** holiday wedding vacation live farm cottage	marry mansion be live castle farm cottage move divorce	honeymoon croft ranch vineyard homestead **residence** farmhouse wedding farm cottage	farm ranch **residence** wedding croft cottage homestead

Table 1: **Examples of generated lexical substitutes**: baselines and our models. Contexts are from the TWSI dataset. Ambiguous word is underlined, substitutes intersecting with human-generated are **bold**. Here **base** is the baseline approach of Amrami and Goldberg (2018) and **base-no-pat** is its simplified version without patterns, while **BComb-LMs** and **BComb-3** are our models described in Section 3.

predicts all words which can appear before dot resulting mostly infrequent abbreviations. Dynamic patterns help a little, but there is still no context available for the backward LM to disambiguate the target word. To solve this problem we propose combining distributions from the forward and the backward LM first and then taking the top K words from this combined distribution. We experiment with the following combinations.

3.2.1 Average (avg)

This straightforward method of fusion of two distributions computes an average of forward and backward distributions (no information about the target word is used):

$$\begin{aligned} P(w|l,r) &= \frac{1}{2}(P(w|l) + P(w|r)) \\ &= \frac{1}{2}(P_{fwd} + P_{bwd}). \end{aligned} \tag{1}$$

3.2.2 Bayesian Combination of LMs (BComb-LMs)

Using Bayes' rule and supposing left and right context are independent given possible substitutes we estimate fused distribution as follows:

$$\begin{aligned} P(w|l,r) &= \frac{P(l,r|w)P(w)}{P(l,r)} \\ &= \frac{P(l|w)P(r|w)P(w)}{P(l,r)} \\ &\propto \frac{P(w|l)P(w|r)}{P(w)}. \end{aligned} \tag{2}$$

The numerator is estimated as $P_{fwd}P_{bwd}$, but pretrained ELMo LMs don't contain frequencies of the words in the vocabulary, so we cannot directly estimate the denominator. Instead we approximate it with Zipf distribution (the vocabulary is sorted by frequency):

$$P(w) \propto \frac{1}{(k + rank(w))^s}, \tag{3}$$

where k and s are hyperparameters: the first is needed to perform adjustment for frequent words while the second defines how quickly word frequency drops as its rank grows.

3.2.3 Three-Way Bayesian Combination (BComb-3)

Substitutes should not only be compatible with context, but also similar to the target word c. Amrami and Goldberg (2018) integrate information

about the target word using dynamic patterns, but here we propose a probabilistic approach of fusion of forward and backward distribution with the information about the target word. Namely, we estimate similarity using a scaled dot product of output embeddings from ELMo:

$$P(w|c) \propto \exp(\frac{emb_w^T emb_c}{Temperature}), \quad (4)$$

where $Temperature$ is a hyperparameter which allows scaling this distribution to fit to the LM distributions. Similarly to *BComb-LMs* and supposing the target word is independent from the context given possible substitutes (which can be interpreted as fixing a particular sense of the target):

$$P(w|l,c,r) \propto \frac{P(w|l)P(w|r)P(w|c)}{P^2(w)}. \quad (5)$$

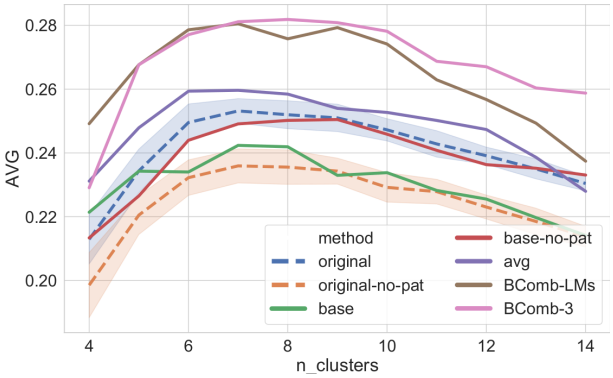

Figure 1: **SemEval 2013 task 13**: geometric average of fNMI and fB^3 with respect to the number of clusters per word. Hyperparameters are selected on the TWSI dataset (Biemann, 2012).

4 Evaluation and Results

To evaluate the quality of our proposed approach, we performed three experiments. Two of them are based on WSI datasets coming from the shared tasks for English (Jurgens and Klapaftis, 2013) and Russian (Panchenko et al., 2018). The last experiment compares substitutes generated by the original and our methods to the human-generated substitutes using a lexical substitution dataset for English (Biemann, 2012).

4.1 Experiment 1: SemEval 2013 WSI Task

4.1.1 Experimental Setup

First, we evaluate our methods on the SemEval-2013 dataset for English WSI (Jurgens and Klapaftis, 2013). The dataset contains contexts for 50 ambiguous words, including 20 nouns, 20 verbs, and 10 adjectives. It provides 20-100 contexts per word, and 4,664 contexts in total, which were gathered from the Open American National Corpus and annotated with senses from WordNet. We used this dataset as the test set and tuned all hyperparameters except for the number of clusters on the TWSI dataset (Biemann, 2012).

Evaluation metrics. Performance is measured with two cluster comparison measures: Fuzzy NMI (fNMI) and Fuzzy B-Cubed (fB^3) as defined in (Jurgens and Klapaftis, 2013).

4.1.2 Discussion of Results

Figure 1 shows a geometric average (AVG) between Fuzzy Normalized Mutual Information (fNMI) and Fuzzy B-Cubed F1 (fB^3) depending on the number of clusters. Following Amrami and Goldberg (2018), Table 2 reports the results for the number of clusters equal to 7 which is the average number of senses in SemEval-2013. BComb-3 shows the best results closely followed by BComb-LMs, while the *avg* combination methods performs worse but still outperforms baseline methods.

4.2 Experiment 2: RUSSE 2018 WSI Task

4.2.1 Experimental Setup

For the Russian language we test our methods on the *active-dict* and the *bts-rnc* datasets from the RUSSE 2018 WSI shared task (Panchenko et al., 2018). These datasets are split into dev and test parts containing non-overlapping ambiguous words. The *bts-rnc* dataset relies on contexts sampled from the Russian National Corpus (RNC)[2] and annotated based on the sense inventory of the Large Explanatory Dictionary of Russian[3]. The dev set contains 30 ambiguous words and 3,491 contexts. The test set contains 51 ambiguous words and 6,556 contexts. The *active-dict* dataset is based on the Active Dictionary of Russian, which is an explanatory dictionary (Apresjan, 2011). For each sense, contexts were extracted from the glosses and examples of this dictionary. The train/development set has 85 ambiguous words and 2,073 contexts. The test set has 168 ambiguous words and 3,729 contexts.

[2] http://ruscorpora.ru/en
[3] http://gramota.ru/slovari/info/bts

Model	fNMI	fB^3	AVG
One sense for all	0.000	**0.623**	0.000
One sense per instance	0.071	0.000	0.000
Best competition results (Jurgens and Klapaftis, 2013)			
AI-KU	0.065	0.390	0.159
Unimelb	0.060	0.483	0.170
Best after-competition results			
(Amrami and Goldberg, 2018)	0.113	0.575	0.254
(Amplayo et al., 2019)	0.096	**0.622**	0.244
This paper			
avg	0.120	0.562	0.260
BComb-LMs	**0.139**	0.566	0.280
BComb-3	0.135	0.586	**0.281**

Table 2: **SemEval 2013 task 13**: comparison to the previous best results. Following Amrami and Goldberg (2018) the number of clusters is 7, other hyperparameters are selected on the TWSI dataset.

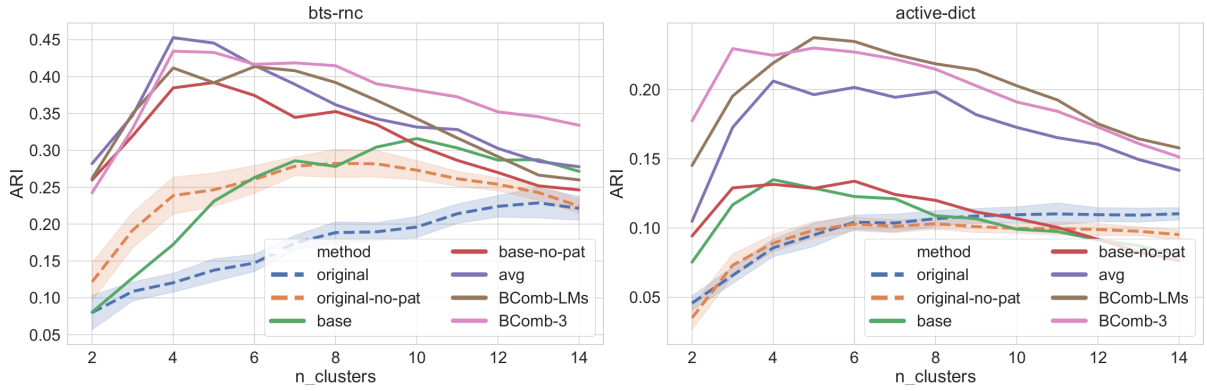

Figure 2: **RUSSE-2018 development sets**: ARI with respect to the number of clusters per word. Hyperparameters are selected on the TWSI dataset.

Evaluation metrics. Performance is measured using Adjusted Rand Index (ARI) (Hubert and Arabie, 1985).

4.2.2 Discussion of Results

Figure 2 shows results on the development set using the same hyperparameters used for SemEval-2013. Despite being selected on an English WSI dataset, they perform surprisingly well. Similarly to SemEval-2013, on *active-dict* BComb methods outperform Avg by a large margin. However, on *bts-rnc* dataset, Avg seems to be the best performing method which we attribute to suboptimal hyperparameters. For our final submissions to the leaderboard reported in Table 3 we selected hyperparameters on the development set corresponding to each dataset and with these hyperparameters BComb methods are indeed better than Avg. We report results for (i) a fixed number of clusters (selected on the development sets) and for (ii) individual number of clusters for each word selected by maximizing the silhouette score of clustering[4]. Using individual number of clusters consistently improves results for all our methods.

4.3 Experiment 3: TWSI Lexical Substitution

4.3.1 Experimental Setup

In the third experiment, we evaluated the quality of lexical substitutes generated by our methods comparing them with human-generated ones from the

[4] https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient

Model	bts-rnc Test	active-dict Test
avg	0.355 / 0.436	0.254 / 0.255
BComb-LMs	0.464 / **0.502**	0.304 / 0.331
BComb-3	0.455 / 0.473	0.300 / **0.332**
post compet'n best results	0.348	0.307
competition 1st best result	0.351	0.264
competition 2nd best result	0.281	0.236

Table 3: **RUSSE 2018 test sets**: comparison to the previous best results. The number of clusters is selected on corresponding development sets (like other hyperparameters) / using silhouette score.

TWSI dataset by Biemann (2012). Version 2.0 of the dataset was used in our experiments. The dataset is composed of 1,012 frequent nouns with 2.26 senses per word on average. For these nouns, the dataset provides 145,140 annotated sentences sampled from Wikipedia. Besides, it features a sense inventory, where each sense is represented with a list of words that can substitutes.

Evaluation Metrics Performance is measured using precision and recall among top $K = 10$ lexical substitutions.

4.3.2 Discussion of Results

Table 4 reports the results. One should carefully interpret these results since humans generate precise but not exhaustive lists of substitutes. For instance, for the sentence *Henry David Thoreau wrote the famous phrase, "In wildness is the preservation of the world."* BComb-3 model generates the following substitutes: *dictum, proverb, poem, motto, epitaph, slogan, quote, aphorism, maxim* from which only *slogan* and *maxim* were generated by humans. As one may observe, according to metrics, both base methods with patterns and BComb-3 generate much more human-like substitutes than their counterparts that do not take into account the target word (base-no-pat and BComb-LMs) with BComb-3 being a little better. Examples of generated substitutes are shown in Table 1.

Model	rec.@10	prec.@10
base	0.115	0.035
base-no-pat	0.058	0.020
avg	0.093	0.032
BComb-LMs	0.073	0.025
BComb-3	**0.127**	**0.041**

Table 4: **TWSI lexical substitution**: comparison our method to the baseline model by Amrami and Goldberg (2018) on the dataset of human-generated lexical substitutes.

5 Conclusion

We proposed a new method for neural word sense induction which improves the approach of Amrami and Goldberg (2018). We show that substantially better results can be obtained if the information from the forward and the backward LMs is combined in a more principled way using Bayesian fusion of distributions rather than a simple union of substitutes generated independently from each distribution. More specifically, this work shows that integration of the forward and the backward distributions retrieved from neural LMs and the similarity to the target word results in better-generated substitutes for ambiguous words, which enabled achieving a new state-of-the-art for WSI for two languages.

Acknowledgements

We thank three anonymous reviewers for their helpful comments and suggestions, especially the Reviewer #2. Besides, we are grateful to all colleagues with whom we discussed our ideas with, especially to Artem Grachev, Dima Lipin, and Alex Nevidomsky.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. As-

sociation for Computational Linguistics, pages 7–12.

Domagoj Alagić, Jan Šnajder, and Sebastian Padó. 2018. Leveraging lexical substitutes for unsupervised word sense induction. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Reinald Kim Amplayo, Seung won Hwang, and Min Song. 2019. Autosense model for word sense induction. In *AAAI*.

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*. Association for Computational Linguistics, Brussels, Belgium, pages 4860–4867. https://www.aclweb.org/anthology/D18-1523.

Saba Anwar, Dmitry Ustalov, Nikolay Arefyev, Simone Paolo Ponzetto, Chris Biemann, and Alexander Panchenko. 2019. HHMM at SemEval-2019 task 2: Unsupervised frame induction using contextualized word embeddings. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, pages 125–129. https://www.aclweb.org/anthology/S19-2018.

Valentina Apresjan. 2011. Active dictionary of the russian language: theory and practice. *Meaning-Text Theory* 2011:13–24.

Nikolay Arefyev, Pavel Ermolaev, and Panchenko Alexander. 2018. Russian word sense induction by clustering averaged word embeddings. In *Proceedings of the 24th International Conference on Computational Linguistics and Intellectual Technologies (Dialogue2018)*. RGGU, Moscow, Russia.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2016. Breaking sticks and ambiguities with adaptive skip-gram. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Osman Baskaya, 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)*. Association for Computational Linguistics, Atlanta, Georgia, USA, pages 300–306. https://www.aclweb.org/anthology/S13-2050.

Chris Biemann. 2006. Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the first workshop on graph based methods for natural language processing*. Association for Computational Linguistics, pages 73–80.

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

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.

Delbert Dueck and Brendan J Frey. 2007. Non-metric affinity propagation for unsupervised image categorization. In *2007 IEEE 11th International Conference on Computer Vision*. IEEE, pages 1–8.

David Hope and Bill Keller. 2013. UoS: A Graph-Based System for Graded Word Sense Induction. 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, 1, pages 689–694. http://www.aclweb.org/anthology/S13-2113.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification* 2(1):193–218.

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)*. Association for Computational Linguistics, Atlanta, Georgia, USA, pages 290–299. https://www.aclweb.org/anthology/S13-2049.

Andrey Kutuzov. 2018. Russian word sense induction by clustering averaged word embeddings. In *Proceedings of the 24th International Conference on Computational Linguistics and Intellectual Technologies (Dialogue2018)*. RGGU, Moscow, Russia.

Jey Han Lau, Paul Cook, and Timothy Baldwin. 2013. unimelb: Topic Modelling-based Word Sense Induction. In *Second Joint Conference on Lexical and Computational Semantics (*SEM): SemEval 2013*. Atlanta, Georgia, USA, volume 2, pages 307–311. http://www.aclweb.org/anthology/S13-2051.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of ICML*. Madison, WI, USA, volume 98, pages 296–304.

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*. Association for Computational Linguistics, Uppsala, Sweden, pages 63–68. https://www.aclweb.org/anthology/S10-1011.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Process-*

ing Systems 26. Curran Associates, Inc., Lake Tahoe, NV, USA, pages 3111–3119.

Roberto Navigli. 2012. 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, pages 115–129.

Alexander Panchenko, Anastasiya Lopukhina, Dmitry Ustalov, Konstantin Lopukhin, Nikolay Arefyev, Alexey Leontyev, and Natalia Loukachevitch. 2018. Russe'2018: a shared task on word sense induction for the russian language. *Computational Linguistics and Intellectual Technologies* pages 547–564.

Ted Pedersen and Rebecca Bruce. 1997. Distinguishing word senses in untagged text. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*. Providence, RI, pages 197–207.

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*. Association for Computational Linguistics, Berlin, Germany, pages 174–183. https://doi.org/10.18653/v1/W16-1620.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, pages 2227–2237. https://doi.org/10.18653/v1/N18-1202.

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*. Los Angeles, California, pages 109–117. http://www.aclweb.org/anthology/N10-1013.

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

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association* 101(476):1566–1581. https://doi.org/10.1198/016214506000000302.

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

Dominic Widdows and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th international conference on Computational linguistics*. Taipei, Taiwan, pages 1–7. https://doi.org/10.3115/1072228.1072342.