A Sense-Topic Model for Word Sense Induction with Unsupervised Data Enrichment

Jing Wang* **Mohit Bansal†** **Kevin Gimpel†** **Brian D. Ziebart*** **Clement T. Yu***
*University of Illinois at Chicago, Chicago, IL, 60607, USA
{jwang69,bziebart,cyu}@uic.edu
†Toyota Technological Institute at Chicago, Chicago, IL, 60637, USA
{mbansal,kgimpel}@ttic.edu

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00122/1566736/tacl_a_00122.pdf by guest on 04 October 2023

Abstract

Word sense induction (WSI) seeks to automatically discover the senses of a word in a corpus via unsupervised methods. We propose a sense-topic model for WSI, which treats sense and topic as two separate latent variables to be inferred jointly. Topics are informed by the entire document, while senses are informed by the local context surrounding the ambiguous word. We also discuss unsupervised ways of enriching the original corpus in order to improve model performance, including using neural word embeddings and external corpora to expand the context of each data instance. We demonstrate significant improvements over the previous state-of-the-art, achieving the best results reported to date on the SemEval-2013 WSI task.

1 Introduction

Word sense induction (WSI) is the task of automatically discovering all senses of an ambiguous word in a corpus. The inputs to WSI are instances of the ambiguous word with its surrounding context. The output is a grouping of these instances into clusters corresponding to the induced senses. WSI is generally conducted as an unsupervised learning task, relying on the assumption that the surrounding context of a word indicates its meaning. Most previous work assumed that each instance is best labeled with a single sense, and therefore, that each instance belongs to exactly one sense cluster. However, recent work (Erk and McCarthy, 2009; Jurgens, 2013) has shown that more than one sense can be used to interpret certain instances, due to context ambiguity and sense relatedness.

To handle these characteristics of WSI (unsupervised, senses represented by token clusters, multiple senses per instance), we consider approaches based on topic models. A topic model is an unsupervised method that discovers the semantic topics underlying a collection of documents. The most popular is latent Dirichlet allocation (LDA; Blei et al., 2003), in which each topic is represented as a multinomial distribution over words, and each document is represented as a multinomial distribution over topics.

One approach would be to run LDA on the instances for an ambiguous word, then simply interpret topics as induced senses (Brody and Lapata, 2009). However, while sense and topic are related, they are distinct linguistic phenomena. Topics are assigned to entire documents and are expressed by *all* word tokens, while senses relate to a single ambiguous word and are expressed through the *local* context of that word. One possible approach would be to only keep the local context of each ambiguous word, discarding the global context. However, the topical information contained in the broader context, though it may not determine the sense directly, might still be useful for narrowing down the likely senses of the ambiguous word.

Consider the ambiguous word *cold*. In the sentence "*His reaction to the experiments was cold*", the possible senses for *cold* include *cold temperature*, *a cold sensation*, *common cold*, or *a negative emotional reaction*. However, if we know that the topic of the document concerns the effects of low temperatures on physical health, then the *negative emotional reaction* sense should become less likely. Therefore, in this case, knowing the topic helps narrow down the set of plausible senses.

Transactions of the Association for Computational Linguistics, vol. 3, pp. 59–71, 2015. Action Editor: Hwee Tou Ng.
Submission batch: 10/2014; Revision batch 12/2014; Revision batch 1/2015; Published 1/2015. ©2015 Association for Computational Linguistics.

At the same time, knowing the sense can also help determine possible topics. Consider a set of texts that all include the word *cold*. Without further information, the texts might discuss any of a number of possible topics. However, if the sense of *cold* is that of *cold ischemia*, then the most probable topics would be those related to organ transplantation.

In this paper, we propose a sense-topic model for WSI, which treats sense and topic as two separate latent variables to be inferred jointly (§4). When relating the sense and topic variables, a bidirectional edge is drawn between them to represent their cyclic dependence (Heckerman et al., 2001). We perform inference using collapsed Gibbs sampling (§4.2), then estimate the sense distribution for each instance as the solution to the WSI task. We conduct experiments on the SemEval-2013 Task 13 WSI dataset, showing improvements over several strong baselines and task systems (§5).

We also present unsupervised ways of enriching our dataset, including using neural word embeddings (Mikolov et al., 2013) and external Web-scale corpora to enrich the context of each data instance or to add more instances (§6). Each data enrichment method gives further gains, resulting in significant improvements over existing state-of-the-art WSI systems. Overall, we find gains of up to 22% relative improvement in fuzzy B-cubed and 50% relative improvement in fuzzy normalized mutual information (Jurgens and Klapaftis, 2013).

2 Background and Related Work

We discuss the WSI task, then discuss several areas of research that are related to our approach, including applications of topic modeling to WSI as well as other approaches that use word embeddings and clustering algorithms.

WSD and WSI: WSI is related to but distinct from word sense disambiguation (WSD). WSD seeks to assign a particular sense label to each target word instance, where the sense labels are known and usually drawn from an existing sense inventory like WordNet (Miller et al., 1990). Although extensive research has been devoted to WSD, WSI may be more useful for downstream tasks. WSD relies on sense inventories whose construction is time-intensive, expensive, and subject to poor inter-annotator agreement (Passonneau et al., 2010). Sense inventories also impose a fixed sense granularity for each ambiguous word, which may not match the ideal granularity for the task of interest. Finally, they may lack domain-specific senses and are difficult to adapt to low-resource domains or languages. In contrast, senses induced by WSI are more likely to represent the task and domain of interest. Researchers in machine translation and information retrieval have found that predefined senses are often not well-suited for these tasks (Voorhees, 1993; Carpuat and Wu, 2005), while induced senses can lead to improved performance (Véronis, 2004; Vickrey et al., 2005; Carpuat and Wu, 2007).

Topic Modeling for WSI: Brody and Lapata (2009) proposed a topic model that uses a weighted combination of separate LDA models based on different feature sets (e.g. word tokens, parts of speech, and dependency relations). They only used smaller units of text surrounding the ambiguous word, discarding the global context of each instance. Yao and Van Durme (2011) proposed a model based on a hierarchical Dirichlet process (HDP; Teh et al., 2006), which has the advantage that it can automatically discover the number of senses. Lau et al. (2012) described a model based on an HDP with positional word features; it formed the basis for their submission (`unimelb`, Lau et al., 2013) to the SemEval-2013 WSI task (Jurgens and Klapaftis, 2013).

Our sense-topic model is distinct from this prior work in that we model sense and topic as two separate latent variables and learn them jointly. We compare to the performance of `unimelb` in §5.

For word sense *disambiguation*, there also exist several approaches that use topic models (Cai et al., 2007; Boyd-Graber and Blei, 2007; Boyd-Graber et al., 2007; Li et al., 2010); space does not permit a full discussion.

Word Representations for WSI: Another approach to solving WSI is to use word representations built by distributional semantic models (DSMs; Sahlgren, 2006) or neural net language models (NNLMs; Bengio et al., 2003; Mnih and Hinton, 2007). Their assumption is that words with similar distributions have similar meanings. Akkaya et al. (2012) use word representations learned from DSMs directly for WSI. Each word is represented by a co-

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00122/1566736/tacl_a_00122.pdf by guest on 04 October 2023

occurrence vector, and the meaning of an ambiguous word in a specific context is computed through element-wise multiplication applied to the vector of the target word and its surrounding words in the context. Then instances are clustered by hierarchical clustering based on their representations.

Word representations trained by NNLMs, often called **word embeddings**, capture information via training criteria based on predicting nearby words. They have been useful as features in many NLP tasks (Turian et al., 2010; Collobert et al., 2011; Dhillon et al., 2012; Hisamoto et al., 2013; Bansal et al., 2014). The similarity between two words can be computed using cosine similarity of their embedding vectors. Word embeddings are often also used to build representations for larger units of text, such as sentences, through vector operations (e.g., summation) applied to the vector of each token in the sentence. In our work, we use word embeddings to compute word similarities (for better modeling of our data distribution), to represent sentences (to find similar sentences in external corpora for data enrichment), and in a product-of-embeddings baseline.

Baskaya et al. (2013) represent the context of each ambiguous word by using the most likely substitutes according to a 4-gram LM. They pair the ambiguous word with likely substitutes, project the pairs onto a sphere (Maron et al., 2010), and obtain final senses via k-means clustering. We compare to their SemEval-2013 system `AI-KU` (§5).

Other Approaches to WSI: Other approaches include clustering algorithms to partition instances of an ambiguous word into sense-based clusters (Schütze, 1998; Pantel and Lin, 2002; Purandare and Pedersen, 2004), or graph-based methods to induce senses (Dorow and Widdows, 2003; Véronis, 2004; Agirre and Soroa, 2007).

3 Problem Setting

In this paper, we induce senses for a set of word types, which we refer to as **target words**. For each target word, we have a set of **instances**. Each instance provides context for a single occurrence of the target word.[1] For our experiments, we use the dataset released for SemEval-2013 Task 13 (Jurgens and Klapaftis, 2013), collected from the Open American National Corpus (OANC; Ide and Suderman, 2004).[2] It includes 50 target words: 20 verbs, 20 nouns, and 10 adjectives. There are a total of 4,664 instances across all target words. Each instance contains only one sentence, with a minimum length of 22 and a maximum length of 100. The gold standard for the dataset was prepared by multiple annotators, where each annotator labeled instances based on the sense inventories in WordNet 3.1. For each instance, they rated all senses of a target word on a Likert scale from one to five.

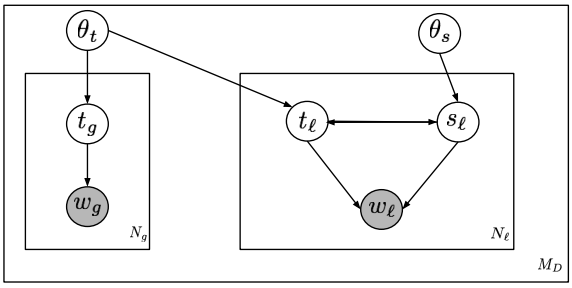

Figure 1: Proposed sense-topic model in plate notation. There are M_D instances for the given target word. In an instance, there are N_g global context words (w_g) and N_ℓ local context words (w_ℓ), all of which are observed. There is one latent variable ("topic" t_g) for the w_g and two latent variables ("topic" t_ℓ and "sense" s_ℓ) for the w_ℓ. Each instance has topic mixing proportions θ_t and sense mixing proportions θ_s. For clarity, not all variables are shown. The complete figure with all variables is given in Appendix A. This is a dependency network, not a directed graphical model, as shown by the directed arrows between t_ℓ and s_ℓ; see text for details.

4 A Sense-Topic Model for WSI

We now present our sense-topic model, shown in plate notation in Figure 1. It generates the words in the set of instances for a single target word; we run the model separately for each target word, sharing no parameters across target words. We treat sense and topic as two separate latent variables to be inferred jointly. To differentiate sense and topic, we use a window around the target word in each instance. Word tokens inside the window are **local**

[1] The target word token may occur multiple times in an instance, but only one occurrence is chosen as the target word occurrence.

[2] "Word Sense Induction for Graded and Non-Graded Senses," `http://www.cs.york.ac.uk/semeval-2013/task13`

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00122/1566736/tacl_a_00122.pdf by guest on 04 October 2023

context words (w_ℓ), while tokens outside the window are **global context words** (w_g). The number of words in the window is fixed to 21 in all experiments (10 words before the target word and 10 after).

Generating global context words: As shown in the left part of Figure 1, each global context word w_g is generated from a latent topic variable t_g for the instance, which follows the same generative story as LDA. The corresponding probability of the ith global context word $w_g^{(i)}$ within instance d is:[3]

$$\Pr(w_g^{(i)}|d,\theta_t,\psi_t) = \sum_{j=1}^{T} P_{\psi_{t_j}}(w_g^{(i)}|t_g^{(i)}=j) P_{\theta_t}(t_g^{(i)}=j|d) \quad (1)$$

where T is the number of topics, $P_{\psi_{t_j}}(w_g^{(i)}|t_g^{(i)}=j)$ is the multinomial distribution over words for topic j (parameterized by ψ_{t_j}) and $P_{\theta_t}(t_g^{(i)}=j|d)$ is the multinomial distribution over topics for instance d (parameterized by θ_t).

Generating local context words: A local context word w_ℓ is generated from a topic variable t_ℓ and a sense variable s_ℓ:

$$\Pr(w_\ell|d,\theta_t,\psi_t,\theta_s,\psi_s,\theta_{s|t},\theta_{t|s},\theta_{st}) = \sum_{j=1}^{T}\sum_{k=1}^{S} \Pr(w_\ell|t_\ell=j,s_\ell=k)\Pr(t_\ell=j,s_\ell=k|d) \quad (2)$$

where S is the number of senses, $\Pr(w_\ell|t_\ell=j,s_\ell=k)$ is the probability of generating word w_ℓ given topic j and sense k, and $\Pr(t_\ell=j,s_\ell=k|d)$ is the joint probability over topics and senses for d.[4]

Unlike in Eq. (1), we do not use multinomial parameterizations for the distributions in Eq. (2). When parameterizing them, we make several departures from purely-generative modeling. All our choices result in distributions over smaller event spaces and/or those that condition on fewer variables. This helps to mitigate data sparsity issues arising from attempting to estimate high-dimensional distributions from small datasets. A secondary benefit is that we can avoid biases caused by particular choices of generative directionality in the model. We later include an empirical comparison to justify some of our modeling choices (§5).

First, when relating the sense and topic variables, we avoid making a single decision about generative dependence. Taking inspiration from dependency networks (Heckerman et al., 2001), we use the following factorization:

$$\Pr(t_\ell=j,s_\ell=k|d) = \frac{1}{Z_d}\Pr(s_\ell=k|d,t_\ell=j)\Pr(t_\ell=j|d,s_\ell=k) \quad (3)$$

where Z_d is a normalization constant.

We factorize further by using redundant probabilistic events, then ignore the normalization constants during learning, a concept commonly called **deficiency** (Brown et al., 1993). Deficient modeling has been found to be useful for a wide range of NLP tasks (Klein and Manning, 2002; May and Knight, 2007; Toutanova and Johnson, 2007). In particular, we factor the conditional probabilities in Eq. (3) into products of multinomial probabilities:

$$\Pr(s_\ell=k|d,t_\ell=j) = \frac{P_{\theta_s}(s_\ell=k|d)P_{\theta_{s|t_j}}(s_\ell=k|t_\ell=j)P_{\theta_{st}}(t_\ell=j,s_\ell=k)}{Z_{d,t_j}}$$

$$\Pr(t_\ell=j|d,s_\ell=k) = \frac{P_{\theta_t}(t_\ell=j|d)P_{\theta_{t|s_k}}(t_\ell=j|s_\ell=k)}{Z_{d,s_k}}$$

where Z_{d,t_j} and Z_{d,s_k} are normalization factors and we have introduced new multinomial parameters $\theta_s, \theta_{s|t_j}, \theta_{st}$, and $\theta_{t|s_k}$.

We use the same idea to factor the word generation distribution:

$$\Pr(w_\ell|t_\ell=j,s_\ell=k) = \frac{P_{\psi_{t_j}}(w_\ell|t_\ell=j)P_{\psi_{s_k}}(w_\ell|s_\ell=k)}{Z_{t_j,s_k}}$$

where Z_{t_j,s_k} is a normalization factor, and we have new multinomial parameters ψ_{s_k} for the sense-word distributions. One advantage of this parameterization is that we naturally tie the topic-word distributions across the global and local context words by using the same parameters ψ_{t_j}.

4.1 Generative Story

We now give the full generative story of our model. We describe it for generating a set of instances of size M_D, where all instances contain the same target word. We use symmetric Dirichlet priors for

[3] We use $\Pr()$ for generic probability distributions without further qualifiers and $P_\theta()$ for distributions parameterized by θ.

[4] For clarity, we drop the (i) superscripts in these and the following equations.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00122/1566736/tacl_a_00122.pdf by guest on 04 October 2023

all multinomial distributions mentioned above, using the same fixed hyperparameter value (α) for all. We use ψ to denote parameters of multinomial distributions over words, and θ to denote parameters of multinomial distributions over topics and/or senses. We leave unspecified the distributions over N_ℓ (number of local words in an instance) and N_g (number of global words in an instance), as we only use our model to perform inference given fixed instances, not to generate new instances.

The generative story first follows the steps described in Algo. 1 to generate parameters that are shared across all instances; then for each instance d, it follows Algo. 2 to generate global and local words.

Algorithm 1 Generative story for instance set

for each topic $j \leftarrow 1$ to T **do**
    Choose topic-word params. $\psi_{t_j} \sim \mathrm{Dir}(\alpha)$
    Choose topic-sense params. $\theta_{s|t_j} \sim \mathrm{Dir}(\alpha)$
for each sense $k \leftarrow 1$ to S **do**
    Choose sense-word params. $\psi_{s_k} \sim \mathrm{Dir}(\alpha)$
    Choose sense-topic params. $\theta_{t|s_k} \sim \mathrm{Dir}(\alpha)$
Choose topic/sense params. $\theta_{st} \sim \mathrm{Dir}(\alpha)$

Algorithm 2 Generative story for instance d

Choose topic proportions $\theta_t \sim \mathrm{Dir}(\alpha)$
Choose sense proportions $\theta_s \sim \mathrm{Dir}(\alpha)$
Choose N_g and N_ℓ from unspecified distributions
for $i \leftarrow 1$ to N_g **do**
    Choose a topic $j \sim \mathrm{Mult}(\theta_t)$
    Choose a word $w_g \sim \mathrm{Mult}(\psi_{t_j})$
for $i \leftarrow 1$ to N_ℓ **do**
    repeat
        Choose a topic $j \sim \mathrm{Mult}(\theta_t)$
        Choose a sense $k \sim \mathrm{Mult}(\theta_s)$
        Choose a topic $j' \sim \mathrm{Mult}(\theta_{t|s_k})$
        Choose a sense $k' \sim \mathrm{Mult}(\theta_{s|t_j})$
        Choose topic/sense $\langle j'', k'' \rangle \sim \mathrm{Mult}(\theta_{st})$
    until $j = j' = j''$ and $k = k' = k''$
    repeat
        Choose a word $w_\ell \sim \mathrm{Mult}(\psi_{t_j})$
        Choose a word $w'_\ell \sim \mathrm{Mult}(\psi_{s_k})$
    until $w_\ell = w'_\ell$

4.2 Inference

We use collapsed Gibbs sampling (Geman and Geman, 1984) to obtain samples from the posterior distribution over latent variables, with all multinomial parameters analytically integrated out before sampling. Then we estimate the sense distribution θ_s for each instance using maximum likelihood estimation on the samples. These sense distributions are the output of our WSI system.

We note that deficient modeling does not ordinarily affect Gibbs sampling when used for computing posteriors over latent variables, as long as parameters (the θ and ψ) are kept fixed. This is the case during the E step of an EM algorithm, which is the usual setting in which deficiency is used. Only the M step is affected; it becomes an approximate M step by assuming the normalization constants equal 1 (Brown et al., 1993).

However, here we use *collapsed* Gibbs sampling for posterior inference, and the analytic integration is disrupted by the presence of the normalization constants. To bypass this, we employ the standard approximation of deficient models that all normalization constants are 1, permitting us to use standard formulas for analytic integration of multinomial parameters with Dirichlet priors. Empirically, we found this "collapsed deficient Gibbs sampler" to slightly outperform a more principled approach based on EM, presumably due to the ability of collapsing to accelerate mixing.

During the sampling process, each sampler is run on the full set of instances for a target word, iterating through all word tokens in each instance. If the current word token is a global context word, we sample a new topic for it conditioned on all other latent variables across instances. If the current word is a local context word, we sample a new topic/sense pair for it again conditioned on all other latent variable values.

We write the conditional posterior distribution over topics for global context word token i in instance d as $\Pr(t_g^{(i)} = j | d, t^{-i}, \boldsymbol{s}, \cdot)$, where $t_g^{(i)} = j$ is the topic assignment of token i, d is the current instance, t^{-i} is the set of topic assignments of all word tokens aside from i for instance d, $\boldsymbol{s}$ is the set of sense assignments for all local word tokens in instance d, and "$\cdot$" stands for all other observed or known information, including all words, all Dirichlet hyperparameters, and all latent variable assignments in other instances. The conditional posterior

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00122/1566736/tacl_a_00122.pdf by guest on 04 October 2023

can be computed by:

$$\Pr(t_g^{(i)} = j | d, t^{-i}, \boldsymbol{s}, \cdot) \propto \underbrace{\frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha}}_{\Pr(t=j|d,t^{-i},\boldsymbol{s},\cdot)} \underbrace{\frac{C_{ij}^{WT} + \alpha}{\sum_{k'=1}^{W_t} C_{k'j}^{WT} + W_t\alpha}}_{\Pr(w_g^{(i)}|t=j,t^{-i},\boldsymbol{s},\cdot)} \quad (4)$$

where we use the superscript DT as a mnemonic for "instance/topic" when counting topic assignments in an instance and WT for "word/topic" when counting topic assignments for a word. C_{dj}^{DT} contains the number of times topic j is assigned to some word token in instance d, excluding the current word token $w_g^{(i)}$; C_{ij}^{WT} is the number of times word $w_g^{(i)}$ is assigned to topic j, across all instances, excluding the current word token. W_t is the number of distinct word types in the full set of instances. We show the corresponding conditional posterior probabilities underneath each term; the count ratios are obtained using standard Dirichlet-multinomial collapsing.

The conditional posterior distribution over topic/sense pairs for a local context word token $w_\ell^{(i)}$ can be computed by:

$$\begin{aligned}
&\Pr(t_\ell^{(i)} = j, s_\ell^{(i)} = k | d, t^{-i}, s^{-i}, \cdot) \propto \\
&\underbrace{\frac{C_{dj}^{DT} + \alpha}{\sum_{k'=1}^{T} C_{dk'}^{DT} + T\alpha}}_{\Pr(t=j|d,t^{-i},\boldsymbol{s},\cdot)} \underbrace{\frac{C_{ij}^{WT} + \alpha}{\sum_{k'=1}^{W_t} C_{k'j}^{WT} + W_t\alpha}}_{\Pr(w_\ell^{(i)}|t=j,t^{-i},\boldsymbol{s},\cdot)} \\
&\underbrace{\frac{C_{dk}^{DS} + \alpha}{\sum_{k'=1}^{S} C_{dk'}^{DS} + S\alpha}}_{\Pr(s=k|d,s^{-i},\cdot)} \underbrace{\frac{C_{ik}^{WS} + \alpha}{\sum_{k'=1}^{W_s} C_{k'k}^{WS} + W_s\alpha}}_{\Pr(w_\ell^{(i)}|s=k,s^{-i},\cdot)} \\
&\underbrace{\frac{C_{kj}^{ST} + \alpha}{\sum_{k'=1}^{S} C_{k'j}^{ST} + S\alpha}}_{\Pr(s=k|t=j,t^{-i},s^{-i},\cdot)} \underbrace{\frac{C_{kj}^{ST} + \alpha}{\sum_{k'=1}^{T} C_{kk'}^{ST} + T\alpha}}_{\Pr(t=j|s=k,t^{-i},s^{-i},\cdot)} \\
&\underbrace{\frac{C_{kj}^{ST} + \alpha}{\sum_{k'=1}^{S}\sum_{j'=1}^{T} C_{k'j'}^{ST} + ST\alpha}}_{\Pr(s=k,t=j|t^{-i},s^{-i},\cdot)}
\end{aligned} \quad (5)$$

where C_{dk}^{DS} contains the number of times sense k is assigned to some local word token in instance d, excluding the current word token; C_{ik}^{WS} contains the number of time word $w_\ell^{(i)}$ is assigned to sense k, excluding the current time; C_{kj}^{ST} contains the number of times sense k and topic j are assigned to some local word tokens. W_s is the number of distinct local context word types across the collection.

Decoding After the sampling process, we obtain a fixed-point estimate of the sense distribution (θ_s) for each instance d using the counts from our samples. Where we use θ_s^k to denote the probability of sense k for the instance, this amounts to:

$$\theta_s^k = \frac{C_{dk}^{DS}}{\sum_{k'=1}^{S} C_{dk'}^{DS}} \quad (6)$$

This distribution is considered the final sense assignment distribution for the target word in instance d for the WSI task; the full distribution is fed to the evaluation metrics defined in the next section.

To inspect what the model learned, we similarly obtain the sense-word distribution (ψ_s) from the counts as follows, where $\psi_{s_k}^i$ is the probability of word type i given sense k:

$$\psi_{s_k}^i = \frac{C_{ik}^{WS}}{\sum_{i'=1}^{W_s} C_{i'k}^{WS}} \quad (7)$$

5 Experimental Results

In this section, we evaluate our sense-topic model and compare it to several strong baselines and state-of-the-art systems.

Evaluation Metrics To evaluate WSI systems, Jurgens and Klapaftis (2013) propose two metrics: fuzzy B-cubed and fuzzy normalized mutual information (NMI). They are each computed separately for each target word, then averaged across target words. Fuzzy B-cubed prefers labeling all instances with the same sense, while fuzzy NMI prefers the opposite extreme of labeling all instances with distinct senses. Hence, we report both fuzzy B-cubed (%) and fuzzy NMI (%) in our evaluation. For ease of comparison, we also report the geometric mean of the 2 metrics, which we denote by AVG.[5]

SemEval-2013 Task 13 also provided a trial dataset (TRIAL) that consists of eight target ambiguous words, each with 50 instances (Erk et al., 2009). We use it for preliminary experiments of our model and for tuning certain hyperparameters, and evaluate final performance on the SemEval-2013 dataset (TEST) with 50 target words.

[5]We do not use an arithmetic mean because the effective range of the two metrics is substantially different.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00122/1566736/tacl_a_00122.pdf by guest on 04 October 2023

S	B-cubed(%)	NMI(%)	AVG
2	**42.9**	4.18	13.39
3	31.9	6.50	**14.40**
5	22.3	8.60	13.85
7	15.4	8.72	11.61
10	12.5	**10.91**	11.67

Table 1: Performance on TRIAL for the sense-topic model with different numbers of senses (S). Best score in each column is bold.

Hyperparameter Tuning We use TRIAL to analyze performance of our sense-topic model under different settings for the numbers of senses (S) and topics (T); see Table 1. We always set $T = 2S$ for simplicity. We find that small S values work best, which is unsurprising considering the relatively small number of instances and small size of each instance. When evaluating on TEST, we use $S = 3$ (which gives the best AVG results on TRIAL). Later, when we add larger context or more instances (see §6), tuning on TRIAL chooses a larger S value.

During inference, the Gibbs sampler was run for 4,000 iterations for each target word, setting the first 500 iterations as the burn-in period. In order to get a representative set of samples, every 13th sample (after burn-in) is saved to prevent correlations among samples. Due to the randomized nature of the inference procedure, all reported results are average scores over 5 runs. The hyperparameters (α) for all Dirichlet priors in our model are set to the (untuned) value of 0.01, following prior work on topic modeling (Griffiths and Steyvers, 2004; Heinrich, 2005).

Baselines We include two naïve baselines corresponding to the two extremes (biases) preferred by fuzzy B-cubed and NMI, respectively: *1 sense* (label each instance with the same single sense) and *all distinct* (label each instance with its own sense).

We also consider two baselines based on LDA. We run LDA for each target word in TEST, using the set of instances as the set of documents. We treat the learned topics as induced senses. When setting the number of topics (senses), we use the *gold-standard* number of senses for each target word, making this *baseline unreasonably strong*. We run LDA both with full context (FULL) and local context (LOCAL), using the same window size as above (10 words before and after the target word).

We also present results for the two best systems in the SemEval-2013 task (according to fuzzy B-cubed and fuzzy NMI, respectively): `unimelb` and `AI-KU`. As described in Section 2, `unimelb` uses hierarchical Dirichlet processes (HDPs). It extracts 50,000 extra instances for each target word as training data from the ukWac corpus−a web corpus of approximately 2 billion tokens.[6] Among all systems in the task, it performs best according to fuzzy B-cubed. `AI-KU` is based on a lexical substitution method; a language model is built to identify lexical substitutes for target words from the dataset and the ukWac corpus. It performed best among all systems according to fuzzy NMI.

Results In Table 2, we present results for these systems and compare them to our basic (i.e., without any data enrichment) sense-topic model with $S = 3$ (row 9). According to both fuzzy B-cubed and fuzzy NMI, our model outperforms the other WSI systems (LDA, `AI-KU`, and `unimelb`). Hence, we are able to achieve state-of-the-art results on the SemEval-2013 task even when only using the single sentence of context given in each instance (while `AI-KU` and `unimelb` use large training sets from ukWac). We found similar performance improvements when only tested on instances labeled with a single sense.

Bidirectionalilty Analysis To measure the impact of the bidirectional dependency between the topic and sense variables in our model, we also evaluate the performance of our sense-topic model when dropping one of the directions. In Table 3, we compare their performance with our full sense-topic model on TEST. Both unidirectional models perform worse than the full model, and dropping $t \rightarrow s$ hurts more. This result verifies our intuition that topics would help narrow down the set of likely senses, and suggests that bidirectional modeling between topic and sense is desirable for WSI.

In subsequent sections, we investigate several ways of exploiting additional data to build better-performing sense-topic models.

6 Unsupervised Data Enrichment

The primary signal used by our model is word co-occurrence information across instances. If we en-

[6] `http://wacky.sslmit.unibo.it/doku.php?id=corpora`

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00122/1566736/tacl_a_00122.pdf by guest on 04 October 2023

	Model	Data Enrichment	Fuzzy B-cubed %	Fuzzy NMI %	AVG
1	1 sense	–	62.3	0	–
2	all distinct	–	0	7.09	–
3	`unimelb`	add 50k instances	48.3	6.0	17.02
4	`AI-KU`	add 20k instances	39.0	6.5	15.92
5	LDA (LOCAL)	none	47.1	5.93	16.71
6	LDA (FULL)	none	47.3	5.79	16.55
7	LDA (FULL)	add actual context (§6.1)	43.5	6.41	16.70
8	word embedding product (§6.3)	none	33.3	7.24	15.53
	THIS PAPER				
9	Sense-Topic Model	none	53.5	6.96	19.30
10		add ukWac context (§6.1)	54.5	**9.74**	23.04
11		add actual context (§6.1)	**59.1**	9.39	**23.56**
12		add instances (§6.2)	58.9	6.01	18.81
13		weight by sim. (§6.3)	55.4	7.14	19.89

Table 2: Performance on TEST for baselines and our sense-topic model. Best score in each column is bold.

Model	B-cubed(%)	NMI(%)	AVG
Drop $s \rightarrow t$	52.1	6.84	18.88
Drop $t \rightarrow s$	51.1	6.78	18.61
Full	53.5	6.96	19.30

Table 3: Performance on TEST for the sense-topic model with ablation of links between sense and topic variables.

rich the instances, we can have more robust co-occurrence statistics. The SemEval-2013 dataset may be too small to induce meaningful senses, since there are only about 100 instances for each target word, and each instance only contains one sentence. This is why most shared task systems added instances from external corpora.

In this section, we consider three unsupervised ways of enriching data and measure their impact on performance. In §6.1 we augment the context of each instance in our original dataset while keeping the number of instances fixed. In §6.2 we collect more instances of each target word from ukWac, similar to the `AI-KU` and `unimelb` systems. In §6.3, we change the distribution of words in each instance based on their similarity to the target word.

Throughout, we make use of word embeddings (see §2). We trained 100-dimensional skip-gram vectors (Mikolov et al., 2013) on English Wikipedia (tokenized/lowercased, resulting in 1.8B tokens of text) using window size 10, hierarchical softmax, and no downsampling.[7]

[7]We used a minimum count cutoff of 20 during training, then only retained vectors for the most frequent 100,000 word types, averaging the rest to get a vector for unknown words.

6.1 Adding Context

The first way we explore of enriching data is to add a broader context for each instance while keeping the number of instances unchanged. This will introduce more word tokens into the set of global context words, while keeping the set of local context words mostly unchanged, as the window size we use is typically smaller than the length of the original instance. With more global context words, the model has more evidence to learn coherent topics, which could also improve the induced senses via the connection between sense and topic.

The ideal way of enriching context for an instance is to add its actual context from the corpus from which it was extracted. To do this for the SemEval-2013 task, we find each instance in the OANC and retrieve three sentences before the instance and three sentences after. While not provided for the SemEval task, it is reasonable to assume this larger context in many real-world applications, such as information retrieval and machine translation of documents.

However, in other settings, the corpus may only have a single sentence containing the target word (e.g., search queries or machine translation of sentences). To address this, we find a semantically-similar sentence from the English ukWac corpus and append it to the instance as additional context. For each instance in the original dataset, we extract its

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00122/1566736/tacl_a_00122.pdf by guest on 04 October 2023

most similar sentence that contains the same target word and add it to increase its set of global context words. To compute similarity, we first represent instances and ukWac sentences by summing the word embeddings across their word tokens, then compute cosine similarity. The ukWac sentence (s^*) with the highest cosine similarity to each original instance (d) is appended to that instance:

$$s^* = \arg\max_{s \in \text{ukWac}} \text{sim}(d, s)$$

Results Since the vocabulary has increased, we expect we may need larger values for S and T. On TRIAL, we find best performance for $S = 10$, so we run on TEST with this value. Performance is shown in Table 2 (rows 10 and 11). These two methods have higher AVG scores than all others. Both their fuzzy B-cubed and NMI improvements over the baselines and previous WSI systems are statistically significant, as measured by a paired bootstrap test ($p < 0.01$; Efron and Tibshirani, 1994).

It is unsurprising that we find best performance with actual context. Interestingly, however, we can achieve almost the same gains when automatically finding relevant context from a *different* corpus. Thus, even in real-world settings where we only have a single sentence of context, we can induce substantially better senses by automatically broadening the global context in an unsupervised manner.

As a comparative experiment, we also evaluate the performance of LDA when adding actual context (Table 2, row 7). Compared with LDA with full context (FULL) in row 6, performance is slightly improved, perhaps due to the fact that longer contexts induce more accurate topics. However, those topics are not necessarily related to senses, which is why LDA with only local context actually performs best among all three LDA models. Thus we see that merely adding context does not necessarily help topic models for WSI. Importantly, since our model includes both sense and topic, we are able to leverage the additional context to learn better topics while also improving the quality of the induced senses, leading to our strongest results.

Examples We present examples to illustrate our sense-topic model's advantage over LDA and the further improvement when adding actual context. Consider instances (1) and (2) below, with target word occurrences in bold:

(1) *Nigeria then sent troops to challenge the coup, evidently to restore the president and repair Nigeria's corrupt **image** abroad.* (image%1:07:01::/4)[8]

(2) *When asked about the Bible's literal account of creation, as opposed to the attractive concept of divine creation, every major Republican presidential candidate—even Bauer—has squirmed, ducked, and tried to steer the discussion back to "faith," "morals," and the general idea that humans "were created in the **image** of God."* (image%1:06:00::/2 image%1:09:02::/4)

Both instances share the common word stem *president*. LDA uses this to put these two instances into the same topic (i.e., sense). In our sense-topic model, *president* is a local context word in instance (1) but a global context word in instance (2). So the effect of sharing words is decreased, and these two instances are assigned to different senses by our model. According to the gold standard, the two instances are annotated with different senses, so our sense-topic model provides the correct prediction.

Next, consider instances (3), (4), and (5):

(3) *I have recently deliberately begun to use variations of "kick ass" and "bites X in the ass" because they are colorful, evocative phrases; because, thanks to South Park, ass references are newly familiar and hilarious and because they don't evoke particularly vivid mental **image** of asses any longer.* (image%1:09:00::/4)

(4) *Also, playing video games that require rapid mental rotation of visual **image** enhances the spatial test scores of boys and girls alike.* (image%1:06:00::/4)

(5) *Practicing and solidifying modes of representation, Piaget emphasized, make it possible for the child to free thought from the here and now; create larger images of reality that take into account past, present, and future; and transform those **image** mentally in the service of logical thinking.* (image%1:09:00::/4)

In the gold standard, instances (3) and (4) have different senses while (3) and (5) have the same sense. However, sharing the local context word "*mental*"

[8]This is the gold standard sense label, where image%1:07:01:: indexes the wordnet senses, and 4 is the score assigned by the annotators.The possible range of a score is [1,5].

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00122/1566736/tacl_a_00122.pdf by guest on 04 October 2023

triggers both LDA and our sense-topic model to assign them to the same sense label with high probability. When augmenting the instances by their real contexts, we have a better understanding about the topics. Instance (3) is about phrase variations, instance (4) is about enhancing boys' spatial skills, while instance (5) discusses the effect of make-believe play for children's development.

When LDA is run with the actual context, it leaves (4) and (5) in the same topic (i.e., sense), while assigning (3) into another topic with high probability. This could be because (4) and (5) both relate to child development, and therefore LDA considers them as sharing the same topic. However, topic is not the same as sense, especially when larger contexts are available. Our sense-topic model built on the actual context makes correct predictions, leaving (3) and (5) into the same sense cluster while labeling (4) with a different sense.

6.2 Adding Instances

We also consider a way to augment our dataset with additional instances from an external corpus. We have no gold standard senses for these instances, so we will not evaluate our model on them; they are merely used to provide richer co-occurrence statistics about the target word so that we can perform better on the instances on which we evaluate.

If we added randomly-chosen instances (containing the target word), we would be concerned that the learned topics and senses may not reflect the distributions of the original instance set. So we only add instances that are semantically similar to instances in our original set (Moore and Lewis, 2010; Chambers and Jurafsky, 2011). Also, to avoid changing the original sense distribution by adding too many instances, we only add a single instance for each original instance. As in §6.1, for each instance in the original dataset, we find the most similar sentence in ukWac for each instance using word embeddings and add it into the dataset. Therefore, the number of instances is doubled, and we use the enriched dataset for our sense-topic model.

Results Similarly to §6.1, on TRIAL, we find best performance for $S = 10$, so we run on TEST with this value. As shown in Table 2 (row 12), this improves fuzzy B-cubed by 5.4%, but fuzzy NMI is lower, making the AVG worse than the original model. A possible reason for this is that the sense distribution in the added instances disturbs that in the original set of instances, even though we picked the most semantically similar ones to add.

6.3 Weighting by Word Similarity

Another approach is inspired by the observation that each local context token is treated equally in terms of its contribution to the sense. However, our intuition is that certain tokens are more indicative than others. Consider the target word *window*. Since *glass* evokes a particular sense of *window*, we would like to weight it more highly than, say, *day*.

To measure word relatedness, we use cosine similarity of word embeddings. We (softly) replicate each local context word according to its exponentiated cosine similarity to the target word.[9] The result is that the local context in each instance has been modified to contain fewer occurrences of unrelated words and more occurrences of related words. If each cosine similarity is 0, we obtain our original sense-topic model. During inference, the posterior sense distribution for instance d is now given by:

$$\Pr(s = k|d, \cdot) = \frac{\sum_{w \in d_\ell} \exp(\text{sim}(w, w^*)) \mathbb{1}_{s_w = k} + \alpha}{\sum_{w' \in d_\ell} \exp(\text{sim}(w', w^*)) + S\alpha} \quad (8)$$

where d_ℓ is the set of local context tokens in d, $\text{sim}(w, w^*)$ is the cosine similarity between w and target word w^*, and $\mathbb{1}_{s_w=k}$ is an indicator returning 1 when w is assigned to sense k and 0 otherwise.

The posterior distribution of sampling a token of word w_i from sense k becomes:

$$\frac{C^{WS}_{ik} \exp(\text{sim}(w_i, w^*)) + \alpha}{\sum_{i'=1}^{W_s} C^{WS}_{i'k} \exp(\text{sim}(w_{i'}, w^*)) + W_s\alpha} \quad (9)$$

where C^{WS}_{ik} counts the number of times w_i is assigned to sense k.

Results We again use TRIAL to tune S (and still use $T = 2S$). We find best TRIAL performance at $S = 3$; this is unsurprising since this approach does not change the vocabulary. In Table 2, we present results on TEST with $S = 3$ (row 13). We also report

[9] Cosine similarities range from -1 to 1, so we use exponentiation to ensure we always use positive counts.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00122/1566736/tacl_a_00122.pdf by guest on 04 October 2023

Sense	Top-5 terms per sense
Sense-Topic Model	
1	*include, depict, party, paint, visual*
2	*zero, manage, company, culture, figure*
3	*create, clinton, people, american, popular*
+weight by similarity (§6.3)	
1	*depict, create, culture, mental, include*
2	*picture, visual, pictorial, matrix, movie*
3	*public, means, view, american, story*

Table 4: Top 5 terms for each sense induced for the noun *image* by the sense-topic model and when weighting local context words by similarity. $S = 3$ for both.

an additional baseline: "word embedding product" (row 8), where we represent each instance by multiplying (element-wise) the word vectors of all local context words, and then feed the instance vectors into the fuzzy c-means clustering algorithm (Pal and Bezdek, 1995), $c = 3$. Compared to this baseline, our approach improves 4.36% on average; compared with results for the original sense-topic model (row 9), this approach improves 0.69% on average.

In Table 4 we show the top-5 terms for each sense induced for *image*, both for the original sense-topic model and when additionally weighting by similarity. We find that the original model provides less distinguishable senses, as it is difficult to derive separate senses from these top terms. In contrast, senses learned from the model with weighted similarities are more distinct. Sense 1 relates to mental representation; sense 2 is about visual representation produced on a surface; and sense 3 is about the general impression that something presents to the public.

7 Conclusions and Future Work

We presented a novel sense-topic model for the problem of word sense induction. We considered sense and topic as distinct latent variables, defining a model that generates global context words using topic variables and local context words using both topic and sense variables. Sense and topic are related using a bidirectional dependency with a robust parameterization based on deficient modeling.

We explored ways of enriching data using word embeddings from neural language models and external corpora. We found enriching context to be most effective, even when the original context of the instance is not available. Evaluating on the SemEval-2013 WSI dataset, we demonstrate that our model yields significant improvements over current state-of-the-art systems, giving 59.1% fuzzy B-cubed and 9.39% fuzzy NMI in our best setting. Moreover, we find that modeling both sense and topic is critical to enable us to effectively exploit broader context, showing that LDA does not improve when each instance is enriched by actual context.

In future work, we plan to further explore the space of sense-topic models, including non-deficient models. One possibility is to use "switching variables" (Paul and Girju, 2009) to choose whether to generate each word from a topic or sense, with a stronger preference to generate from senses closer to the target word. Another possibility is to use locally-normalized log-linear distributions and include features pairing words with particular senses and topics, rather than redundant generative steps.

Appendix A

The plate diagram for the complete sense-topic model is shown in Figure 2.

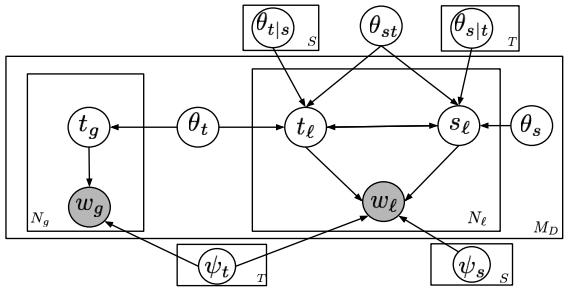

Figure 2: Plate notation for the proposed sense-topic model with all variables (except α, the fixed Dirichlet hyperparameter used as prior for all multinomial distributions). Each instance has topic mixing proportions θ_t and sense mixing proportions θ_s. The instance set shares sense/topic parameter θ_{st}, topic-sense distribution $\theta_{s|t}$, sense-topic distribution $\theta_{t|s}$, topic-word distribution ψ_t, and sense-word distribution ψ_s.

Acknowledgments

We thank the editor and the anonymous reviewers for their helpful comments. This research was partially supported by NIH LM010817. The opinions expressed in this work are those of the authors and do not necessarily reflect the views of the funding agency.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00122/1566736/tacl_a_00122.pdf by guest on 04 October 2023

References

E. Agirre and A. Soroa. 2007. SemEval-2007 Task 02: Evaluating word sense induction and discrimination systems. In *Proc. of SemEval*, pages 7–12.

C. Akkaya, J. Wiebe, and R. Mihalcea. 2012. Utilizing semantic composition in distributional semantic models for word sense discrimination and word sense disambiguation. In *Proc. of ICSC*, pages 45–51.

M. Bansal, K. Gimpel, and K. Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proc. of ACL*, pages 809–815.

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

Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

J. Boyd-Graber and D. M. Blei. 2007. PUTOP: Turning predominant senses into a topic model for word sense disambiguation. In *Proc. of SemEval*, pages 277–281.

J. Boyd-Graber, D. M. Blei, and X. Zhu. 2007. A topic model for word sense disambiguation. In *Proc. of EMNLP-CoNLL*, pages 1024–1033.

S. Brody and M. Lapata. 2009. Bayesian word sense induction. In *Proc. of EACL*, pages 103–111.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

J. F. Cai, W. S. Lee, and Y. W. Teh. 2007. Improving word sense disambiguation using topic features. In *Proc. of EMNLP-CoNLL*, pages 1015–1023.

M. Carpuat and D. Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *Proc. of ACL*, pages 387–394.

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

N. Chambers and D. Jurafsky. 2011. Template-based information extraction without the templates. In *Proc. of ACL*, pages 976–986.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

P. Dhillon, J. Rodu, D. Foster, and L. Ungar. 2012. Two Step CCA: A new spectral method for estimating vector models of words. In *ICML*, pages 1551–1558.

B. Dorow and D. Widdows. 2003. Discovering corpus-specific word senses. In *Proc. of EACL*, pages 79–82.

B. Efron and R. J. Tibshirani. 1994. *An introduction to the bootstrap*, volume 57. CRC press.

K. Erk and D. McCarthy. 2009. Graded word sense assignment. In *Proc. of EMNLP*, pages 440–449.

K. Erk, D. McCarthy, and N. Gaylord. 2009. Investigations on word senses and word usages. In *Proc. of ACL*, pages 10–18.

S. Geman and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741.

T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proc. of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235.

D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. 2001. Dependency networks for inference, collaborative filtering, and data visualization. *J. Mach. Learn. Res.*, 1:49–75.

G. Heinrich. 2005. Parameter estimation for text analysis. Technical report.

S. Hisamoto, K. Duh, and Y. Matsumoto. 2013. An empirical investigation of word representations for parsing the web. In *ANLP*.

N. Ide and K. Suderman. 2004. The American National Corpus first release. In *Proc. of LREC*, pages 1681–1684.

D. Jurgens and I. Klapaftis. 2013. SemEval-2013 Task 13: Word sense induction for graded and non-graded senses. In *Proc. of SemEval*, pages 290–299.

D. Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proc. of NAACL*, pages 556–562.

D. Klein and C. D. Manning. 2002. A generative constituent-context model for improved grammar induction. In *Proc. of ACL*, pages 128–135.

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

J. H. Lau, P. Cook, and T. Baldwin. 2013. unimelb: Topic modelling-based word sense induction. In *Proc. of SemEval*, pages 307–311.

L. Li, B. Roth, and C. Sporleder. 2010. Topic models for word sense disambiguation and token-based idiom detection. In *Proc. of ACL*, pages 1138–1147.

Y. Maron, E. Bienenstock, and M. James. 2010. Sphere embedding: An application to part-of-speech induction. In *Advances in NIPS 23*.

J. May and K. Knight. 2007. Syntactic re-alignment models for machine translation. In *Proc. of EMNLP-CoNLL*, pages 360–368.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00122/1566736/tacl_a_00122.pdf by guest on 04 October 2023

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *Proc. of ICLR*.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4).

A. Mnih and G. Hinton. 2007. Three new graphical models for statistical language modelling. In *Proc. of ICML*, pages 641–648.

R. C. Moore and W. Lewis. 2010. Intelligent selection of language model training data. In *Proc. of ACL*, pages 220–224.

N. R. Pal and J. C. Bezdek. 1995. On cluster validity for the fuzzy c-means model. *Trans. Fuz Sys.*, 3:370–379.

P. Pantel and D. Lin. 2002. Discovering word senses from text. In *Proc. of KDD*, pages 613–619.

R. J. Passonneau, A. Salleb-Aoussi, V. Bhardwaj, and N. Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In *Proc. of LREC*.

M. Paul and R. Girju. 2009. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proc. of EMNLP*, pages 1408–1417.

A. Purandare and T. Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proc. of CoNLL*, pages 41–48.

M. Sahlgren. 2006. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. *Ph.D. dissertation, Stockholm University*.

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

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.

K. Toutanova and M. Johnson. 2007. A Bayesian LDA-based model for semi-supervised part-of-speech tagging. In *Advances in NIPS 20*.

J. Turian, L. Ratinov, and Y. Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proc. of ACL*, pages 384–394.

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

D. Vickrey, L. Biewald, M. Teyssier, and D. Koller. 2005. Word-sense disambiguation for machine translation. In *Proc. of HLT-EMNLP*, pages 771–778.

E. M. Voorhees. 1993. Using WordNet to disambiguate word senses for text retrieval. In *Proc. of SIGIR*, pages 171–180.

X. Yao and B. Van Durme. 2011. Nonparametric Bayesian word sense induction. In *Proc. of TextGraphs-6: Graph-based Methods for Natural Language Processing*, pages 10–14.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00122/1566736/tacl_a_00122.pdf by guest on 04 October 2023

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00122/1566736/tacl_a_00122.pdf by guest on 04 October 2023