Granularity-Agnostic Sense Model for Word Sense Induction

Reinald Kim Amplayo* **Seung-won Hwang†** **Min Song†**
*University of Edinburgh †Yonsei University
reinald.kim@ed.ac.uk {seungwonh, min.song}@yonsei.ac.kr

Abstract

Word sense induction (WSI), or the task of automatically discovering multiple senses or meanings of a word, has three main challenges: domain adaptability, novel sense detection, and sense granularity flexibility. While current latent variable models are known to solve the first two challenges, they are not flexible to different word sense granularities, which differ very much among words, from *aardvark* with one sense, to *play* with over 50 senses. Current models either require hyperparameter tuning or automatically inducing the number of senses, which we find both to be ineffective. Thus, we aim to eliminate these requirements and solve the **sense granularity problem** by proposing **Granularity-Agnostic Sense Model (GAS)**, a latent variable model based on two observations: (1) senses are represented as a distribution over topics, and (2) senses generate pairings between the target word and its neighboring word. These observations alleviate the problem by (a) throwing garbage senses and (b) additionally inducing fine-grained word senses. Results show great improvements over the state-of-the-art models on popular WSI datasets. We also show that GAS is able to learn the appropriate sense granularity of a word. Finally, we apply GAS to the unsupervised author name disambiguation task where the sense granularity problem is more evident and show that GAS is evidently better than competing models. We share our data and code here: http://anonymous.link.

Introduction

Word sense induction (WSI) is the task where given an ambiguous target word (e.g. *cold*) and texts where the word is used, we automatically discover its multiple senses or meanings (e.g. (1) *nose infection*, (2) *absence of heat*, etc.). We show examples of words with multiple senses and example usage in a text[1] in Figure 1. It is distinct from its similar supervised counterpart, word sense disambiguation (WSD) (Stevenson and Wilks 2003), because WSI models should consider the following challenges due to its unsupervised nature: (C1) adaptability to new domains, (C2) ability to detect novel senses, and (C3) flexibility to different word sense granularities (Jurgens and Klapaftis 2013). Another task similar to the WSI is the unsupervised author name disambiguation (UAND) task (Song et al. 2007), where it aims to automatically find different authors, instead of words, with the same name.

Senses of **cold**

- S: (n) **cold**, common cold (a mild viral infection involving the nose and respiratory passages (but not the lungs)) *"will they never find a cure for the common cold?"*
- S: (n) coldness, **cold**, low temperature, frigidity, frigidness (the absence of heat) *"the coldness made our breath visible"; "come in out of the cold"; "cold is a vasoconstrictor"*
- S: (n) **cold**, coldness (the sensation produced by low temperatures) *"he shivered from the cold"; "the cold helped clear his head"*

Senses of **play**

- S: (n) **play**, drama, dramatic play (a dramatic work intended for performance by actors on a stage) *"he wrote several plays but only one was produced on Broadway"*
- S: (n) **play** (a theatrical performance of a drama) *"the play lasted two hours"*
- S: (n) **play** (a preset plan of action in team sports) *"the coach drew up the plays for her team"*
- S: (n) maneuver, manoeuvre, **play** (a deliberate coordinated movement requiring dexterity and skill) *"he made a great maneuver"; "the runner was out on a play by the shortstop"*
- S: (n) **play** (a state in which action is feasible) *"the ball was still in play"; "insiders said the company's stock was in play"*
- S: (n) **play** (utilization or exercise) *"the play of the imagination"*

Figure 1: Three senses of the noun *cold* and six of 17 senses of the noun *play* in WordNet. Sense granularity problem refers to the inflexibility of the model to the different number of senses different words may have (i.e. 3 vs. 17).

In this paper, we consider a latent variable modeling approach to WSI problem as it is proven to be more effective than other approaches (Chang, Pei, and Chen 2014; Komninos and Manandhar 2016). Specifically, we look into methods based on Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), a topic modeling method that automatically discovers the topics underlying a set of documents using Dirichlet priors to infer the multinomial distribution over words and topics. LDA naturally answers two of the three main problems mentioned above, i.e. (C1) and (C2), of the WSI task (Brody and Lapata 2009). However, it is not flexible with regards to (C3), or the **sense granularity problem**, as it requires the users to specify the number of senses: Current systems (Wang et al. 2015; Chang, Pei, and Chen 2014) required to set the number of senses to a small number (set to 3 or 5 in the literature) to get a good accuracy, however many words may have a large number of senses, e.g. *play* in Figure 1.

[1]All sense meanings are copied from WordNet: http://wordnetweb.princeton.edu/perl/webwn

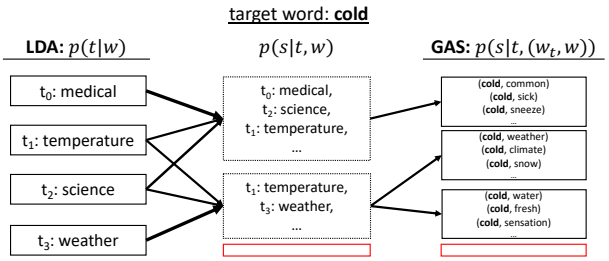


Figure 2: Example induced senses when the target word is *cold* from LDA and GAS. Applying our observations to LDA introduces both garbage and fine-grained senses.

To this end, we propose a latent variable model called **Granularity-Agnostic Sense Model (GAS)**, that solves all the challenges of WSI, including overcoming the sense granularity problem. Consider Figure 2 on finding the senses of the target word *cold*. An LDA model naively considers the topics as senses and thus differentiates the usage of *cold* in the *medical* and *science* domains, even though the same sense is commonly used in the two domains. This results in too many senses induced by the model. We extend LDA using two observations. First, we introduce a separate latent variable for senses, which can be represented as a distribution over topics. This introduces more accurate induced senses (e.g. the *cold: nose infection* sense can be from a mixture of medical, science, and temperature topics), as well as **garbage senses** (colored red in the figure) as most topic distributions will not be assigned to any instance. Second, we enforce senses to generate target-neighbor pairs, a pair (w_t, w) which consists of the target word w_t and one of its neighboring word w, at once. This separates the topic distributions into **fine-grained senses** based on lexical semantic features easily captured by the target-neighbor pairs. For example, the *cold: absence of heat* and the *cold: sensation from low temperature* senses are both related to temperature, but have different syntactic and semantic usage.

By applying the two observations above, GAS removes the strict requirement on correctly setting the number of senses by throwing garbage senses and introducing fine-grained senses. Nonparametric models (Teh et al. 2004; Lau, Cook, and Baldwin 2013) have also been used to solve this problem by automatically inducing the number of senses, however our experiments show that these models are less effective than parametric models and induce incorrect number of senses. Our proposed model is parametric, and is also able to adapt to the different number of senses of different words, even when the number of senses is set to an arbitrarily large number. Moreover, the model can also be used in other tasks such as UAND where the variance in the number of senses is large. To the best of our knowledge, we are the first to experiment extensively on the sense granularity problem of parametric latent variable models.

In our experiments, we estimate the parameters of the model using collapsed Gibbs sampling and get the sense distribution of each instance as the WSI solution. We evaluate our model using the SemEval 2010 and 2013 WSI datasets (Manandhar et al. 2010; Jurgens and Klapaftis 2013). Results show that GAS performs superior than previous state-of-the-art models. We also provide analyses and experiments that shows how GAS overcomes the issue on sense granularity. Finally, we show that our model performs the best on unsupervised author name disambiguation (UAND), where the sense granularities are extremely varied.

Related Work

Previous works on WSI used context vectors and attributes (Almuhareb, Poesio, and others 2006), pretrained classification systems (Tsvetkov et al. 2014), and alignment of parallel corpus (Yao, Van Durme, and Callison-Burch 2012). In the most recent shared task on WSI (Jurgens and Klapaftis 2013), top models used lexical substitution method (**AI-KU**) (Baskaya et al. 2013) and Hierarchical Dirichlet Process trained with additional instances (**Unimelb**) (Lau, Cook, and Baldwin 2013).

Latent variable models such as LDA (Blei, Ng, and Jordan 2003) are used to induce the word sense of a target word after rigorous preprocessing and feature extraction (**LDA**, **Spectral**) (Goyal and Hovy 2014). More recent models introduced a latent variable for the sense of a word, with the assumption that a sense has multiple concepts (**HC**, **HC+Zipf**) (Chang, Pei, and Chen 2014) and that topics and senses should be inferred jointly (**STM**) (Wang et al. 2015). In this paper, we also use a separate sense latent variable, however we show boost in performance by representing it with more versatility and by incorporating the use of target-neighbor pairs. HC was also extended to a nonparametric model (**BNP-HC**) (Teh et al. 2004) in order to automatically set the number of senses of a word, providing flexibility to the sense granularity (Yao and Van Durme 2011; Lau et al. 2012; Lau, Cook, and Baldwin 2013). In our experiments, we show that the sense granularity induced from nonparametric models are incorrect making the models less effective.

Recent inclusions to the WSI models are neural-based dense distributional representation models. STM also used word embeddings (Mikolov et al. 2013) to assign similarity weights during inference (**STM+w2v**) (Wang et al. 2015). Existing sense embeddings are also used to perform word sense induction (**CRP-PPMI**, **SE-WSI-fix**, **WG**, **DIVE**) (Song 2016; Pelevina et al. 2016; Chang et al. 2018). These models, on their own, do not perform well on the WSI task until recently when embeddings of words and their dependencies are used to construct a probabilistic model (**MCC**) (Komninos and Manandhar 2016). We show that neural-based embeddings are still ineffective for this task and that our model performs better than these models as well.

In the unsupervised author name disambiguation (UAND) domain, LDA-based models have also been used (Shu, Long, and Meng 2009) to employ text features for the task, while non-text features such as co-authors, publication venue, year, and citations are found to be stronger features (Tang et al. 2012). In this paper, we study on how to improve the performance of text features for UAND using latent variable models, which can later be combined with non-text features in the future work.

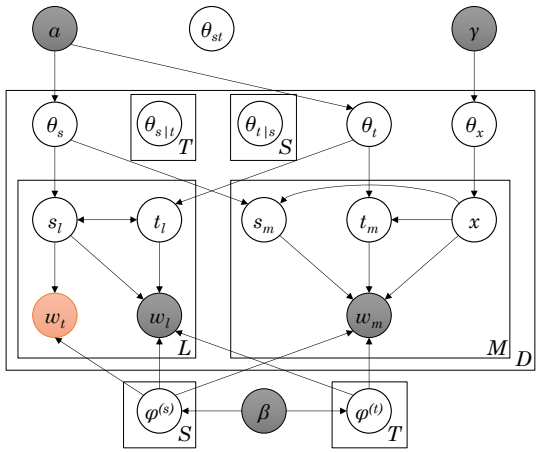


Figure 3: Graphical representation of GAS. Nodes are random variables, edges are dependencies, and plates are replications. Nodes shaded in black are observed. The node shaded in red is the observed target word. The dependency edges of $\theta_{s|t}$, $\theta_{t|s}$, and θ_{st} are not shown for clarity: They are all generated by the Dirichlet prior α. Moreover, sense variables are dependent to $\theta_{s|t}$ and θ_{st}, while topic variables are dependent to $\theta_{t|s}$ and θ_{st}.

D	# of documents
L	# of local context words
M	# of global context words
S	# of senses
T	# of topics
V	vocabulary size
w_t	target word
w_l, w_m	word in local/global context
s_l, s_m	sense in local/global context
t_l, t_m	topic in local/global context
x	sense/topic switch
$\theta_s, \theta_t, \theta_x$	multinomial distribution over senses/topics/switches
$\theta_{s\|t}, \theta_{t\|s}$	multinomial distribution over senses/topics given topics/senses
θ_{st}	multinomial distribution over sense & topic pairs
$\phi^{(s)}, \phi^{(t)}$	multinomial distribution over words
α	Dirichlet prior over θs, except θ_x
β	Dirichlet prior over ϕs
γ	Dirichlet prior over θ_x

Table 1: Meanings of the notations in GAS

Proposed Model

There are two reasons why Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) is not effective for WSI. First, LDA tries to give instance assignments to all senses even when it is unnecessary. For example, when the number of senses S is set to 10, the model tries to assign all the senses to all instances even when the original number of senses of a target word is 3. LDA extensions (Wang et al. 2015; Chang, Pei, and Chen 2014) mitigated this problem by setting S to a small number (e.g. 3 or 5). However, this is not a good solution because there are many words with more than five senses. Second, LDA and its extensions do not consider the existence of fine-grained senses. For example, the *cold: absence of heat* and the *cold: sensation from low temperature* senses are fine-grained senses because they are similarly related to temperature yet have different usage.

Granularity-Agnostic Sense Model

To solve the problems above, we propose to extend LDA in two parts. First, we introduce a new latent variable, apart from the topic latent variable, to represent word senses. Previous works also attempted to introduce a separate sense latent variable to generate all the words (Chang, Pei, and Chen 2014), or to generate only the neighboring words within a local context, decided by a strict user-specified window (Wang et al. 2015). We improve by softening the strict local context assumption by introducing a switch variable which decides whether a word not in a local context should be generated by conditioning also on the sense latent variable. Our experiments show that our sense representation provides superior improvements from previous models.

Second, we force the model to generate target-neighbor pairs at once in the local context, instead of generating words one by one. A target-neighbor pair (w_t, w) consists of the target word w_t and a neighboring word w in the local context. For example, the target-neighbor pairs in "*cold* snowy weather", where w_t is *cold*, are $(cold, snowy)$ and $(cold, weather)$. These pairs give explicit information on the lexical semantics of the target word given the neighboring words. In our running example (Figure 2), the *cold: absence of heat* and the *cold: sensation from low temperature* senses can be easily differentiated when we are given the target-neighbor pairs $(cold, weather)$ and $(cold, climate)$ for the former, and $(cold, water)$ and $(cold, fresh)$ for the latter sense, rather than the individual words.

These extensions bring us to our proposed model called **Granularity-Agnostic Sense Model (GAS)**. The graphical representation of GAS is shown in Figure 3, while the meaning of the notations used in this paper is shown in Table 1.

Generative process For each instance, we divide the text into two contexts: the local context L which includes the target word w_t and its neighboring words w_l, and the global context M which contains the other remaining words w_m. Words from different contexts are generated separately.

In the global context M, words w_m are generated from either a sense s or a topic t latent variable. The selection is done by a switch variable x. If $x = 1$, then the word generation is done by using the sense variable s. Otherwise, it is done by using the topic variable t. The probability of a global context word w_m in document d is given below.

$$\begin{aligned} P(w_m|d) = \; & P(w_m|x=1)\sum_s P(w_m|s)P(s|d) + \\ & P(w_m|x=2)\sum_t P(w_m|t)P(t|d) \\ = \; & \theta_{x=1}\sum_s \theta_s^{(d)}\phi_{w_m}^{(s)} + \theta_{x=2}\sum_t \theta_t^{(d)}\phi_{w_m}^{(t)} \end{aligned}$$

In the local context L, words w_l are generated from both sense s and topic t variables. Also, the target word w_t is generated along with w_l as target-neighbor pairs (w_t, w_l)

using the sense variable s. Sense and topic variables are dependent to each other, so we generate them using the joint probability $p(s,t|d)$. We factorize $p(s,t|d)$ approximately using ideas from dependency networks (Heckerman et al. 2000) to avoid independency assumptions, i.e. $p(a,b|c) = p(a|b,c)p(b|a,c)$, and deficient modeling (Brown et al. 1993) to ignore redundancies, i.e. $p(a|b,c)p(b|a,c) = p(a|b)p(a|c)p(b|a)p(b|c)p(a,b)$. The probability of a local context word w_l in document d given below.

$$\begin{aligned}
P(w_t, w_l|d) &= \sum_s \sum_t \quad p(w_t|s)p(w_l|s,t)p(s,t|d) \\
&\approx \sum_s \sum_t \quad p(w_t|s)p(w_l|s,t)p(s|d,t)p(t|d,s) \\
&\approx \sum_s \sum_t \quad p(w_t|s)p(w_l|s)p(w_l|t) \\
&\qquad\qquad p(s|d)p(s|t)p(t|d)p(t|s)p(s,t) \\
&= \sum_s \sum_t \quad \phi^{(s)}_{w_t}\phi^{(s)}_{w_l}\phi^{(t)}_{w_l}\theta^{(d)}_s\theta^{(d)}_t\theta_{s|t}\theta_{t|s}\theta_{st}
\end{aligned}$$

Inference We use collapsed Gibbs sampling (Griffiths and Steyvers 2004) to estimate the latent variables. At each transition step of the Markov chain, for each word w_m in the global context, we draw the switch $x \sim \{1,2\}$, and the sense $s = k$ or the topic $t = j$ variables using the conditional probabilities given below. The variable C^{AB}_{ab} represents the number of $a \in A$ and $b \in B$ assignments, excluding the current word. The $rest$ corresponds to the other remaining variables, such as the instance d, the current word w_m, the θ and ϕ distributions, and the α, β, and γ Dirichlet priors.

$$P(x=1, s=k|rest) = \frac{C^{DX}_{d1} + \gamma}{\sum_{x'=1}^{2} C^{DX}_{dx'} + 2\gamma} \frac{C^{DS}_{dk} + \alpha}{\sum_{k'=1}^{S} C^{DS}_{dk'} + S\alpha} \frac{C^{SW}_{kw_m} + \beta}{\sum_{w'=1}^{V} C^{SW}_{kw'} + V_m\beta}$$

$$P(x=2, t=j|rest) = \frac{C^{DX}_{d2} + \gamma}{\sum_{x'=1}^{2} C^{DX}_{dx'} + 2\gamma} \frac{C^{DT}_{dj} + \alpha}{\sum_{j'=1}^{T} C^{DT}_{dj'} + T\alpha} \frac{C^{TW}_{jw_m} + \beta}{\sum_{w'=1}^{V} C^{TW}_{jw'} + V_m\beta}$$

Subsequently, for each word w_l and the target word w_t (forming the target-neighbor pair (w_t, w_l)) in the local context, we draw the sense $s = k$ and the topic $t = j$ variables using the conditional probability given below.

$$\begin{aligned}
P(t_i = j, s_i = k|rest) = {} & \frac{C^{DT}_{dj} + \alpha}{\sum_{j'=1}^{T} C^{DT}_{dj'} + T\alpha} \frac{C^{DS}_{dk} + \alpha}{\sum_{k'=1}^{S} C^{DS}_{dk'} + S\alpha} \frac{C^{TW}_{jw_l} + \beta}{\sum_{w'=1}^{V} C^{TW}_{jw'} + V_l\beta} \\
& \frac{C^{SW}_{kw_l} + \beta}{\sum_{w'=1}^{V} C^{SW}_{kw'} + V_l\beta} \frac{C^{SW}_{kw_t} + \beta}{\sum_{w'=1}^{V} C^{SW}_{kw'} + V_l\beta + 1} \\
& \frac{C^{ST}_{kj} + \alpha}{\sum_{j'=1}^{T} C^{ST}_{kj'} + T\alpha} \frac{C^{TS}_{jk} + \alpha}{\sum_{k'=1}^{S} C^{TS}_{jk'} + S\alpha} \\
& \frac{C^{ST}_{kj} + \alpha}{\sum_{k'=1}^{S}\sum_{j'=1}^{T} C^{ST}_{k'j'} + ST\alpha_{j'}}
\end{aligned}$$

Word sense induction After inference is done, the approximate probability of the sense s of the target word in a given document d is induced using the sense distribution of the document as shown in the equation below, where C^{AB}_{ab} represents the number of $a \in A$ and $b \in B$ assignments. We also calculate the word distribution of each sense using the second equation below to inspect the meaning of sense.

$$\theta_{s|d} = \frac{C^{DS}_{ds}}{\sum_{s'=1}^{S} C^{DS}_{ds'}} \quad \theta_{w|s} = \frac{C^{SW}_{sw}}{\sum_{w'=1}^{V} C^{SW}_{sw'}} \tag{1}$$

Experimental setup

Datasets and preprocessing We use two publicly available datasets: SemEval 2010 Task 14 (Manandhar et al. 2010) and SemEval 2013 Task 13 (Jurgens and Klapaftis 2013). The SemEval 2010 dataset[2] consists of 50 verbs and 50 nouns, each with different number of instances for a total of 8915 instances. SemEval 2013 dataset[3] consists of 20 verbs, 20 nouns, and 10 adjectives, with a total of 4664 instances.

For preprocessing, we do tokenization, lemmatization, and removing of symbols to build the word lists using Stanford CoreNLP (Manning et al. 2014). We also divide the word lists into two contexts: the local and global context. Following (Wang et al. 2015), we set the local context window to 10, with a maximum number of words of 21 (i.e. 10 words before and 10 words after). Other words are put into the global context. Note however that GAS has a less strict global/local context assumption as it treats some words in the global context as local depending on the switch variable.

Parameter setting We set the hyperparameters to $\alpha = 0.1$, $\beta = 0.01$, $\gamma = 0.3$, following the conventional setup (Griffiths and Steyvers 2004; Chemudugunta, Smyth, and Steyvers 2006). We arbitrarily set the number of senses to $S = 15$, and the number of topics $T = 2S = 30$, following (Wang et al. 2015). We also include four other versions of our model: **GAS**$^{-wp}$ removes the target-neighbor pair constraint and transforms the local context to that of STM, **GAS**$^{-sw}$ removes the switch variable and transforms the global context to that of LDA, **GAS**$^{s=X}$ is a tuned and best version of the model, where the number of senses is tuned over a separate development set provided by the shared tasks and X is the tuned number of sense, different for each dataset, and **GAS**$^{s=100}$ is the overestimated and worst version of the model, where we set the number of senses to an arbitrary large number, i.e. 100.

We set the number of iterations to 2000 and run the Gibbs sampler. Following the convention of previous works (Lau et al. 2012; Goyal and Hovy 2014; Wang et al. 2015), we assume convergence when the number of iterations is high. However, due to the randomized nature of Gibbs sampling, we report the average scores over 5 runs of Gibbs sampling. We then use the distribution $\theta_{s|d}$ as shown in Equation 1 as the solution of the WSI problem.

[2] https://www.cs.york.ac.uk/semeval2010_WSI

[3] https://www.cs.york.ac.uk/semeval-2013/task13/

Model		F-S	V-M	AVG	$\delta(\#S)$
LVMs	LDA	60.7	4.4	16.34	1.40
	Spectral	61.5	4.5	16.64	1.98
	HC	44.4	11.5	22.62	1.15
	HC+Zipf	35.1	15.2	23.10	3.81
	BNP-HC	23.1	**21.4**	22.23	11.77
NBEs	CRP-PPMI	57.7	2.9	12.94	2.09
	SE-WSI-fix	55.1	9.8	23.24	1.35
Ours	GAS^{-wp}	59.3	9.2	23.36	2.16
	GAS^{-sw}	61.1	8.6	22.92	1.42
	GAS	61.7	9.8	24.59	0.33
	GAS$^{s=5}$	**62.9**	10.1	**25.20**	**0.32**
	GAS$^{s=100}$	61.2	9.6	24.23	0.78

(a) SemEval 2010 WSI dataset

Model		F-BC	F-NMI	AVG
Substitution	AI-KU	39.0	6.50	15.92
LVMs	Unimelb	48.3	6.00	17.02
	STM	53.5	6.96	19.30
NBEs	WG	58.1	1.60	9.64
	DIVE	49.9	3.50	13.22
LVMs + NBEs	STM+w2v	55.4	7.14	19.89
	MCC	55.6	7.62	20.58
Ours	GAS^{-wp}	55.7	7.69	20.69
	GAS^{-sw}	61.4	7.36	21.26
	GAS	**61.7**	7.96	22.16
	GAS$^{s=7}$	**61.7**	**7.97**	**22.17**
	GAS$^{s=100}$	61.0	7.25	21.03
with additional contexts				
STM	+actual	59.1	9.39	23.56
	+ukWac	54.5	**9.74**	23.04
GAS	+actual	**62.2**	9.55	**24.37**

(b) SemEval 2013 WSI dataset

Table 2: Performance of different models on the datasets. Best scores are bold-faced. LVMs are Latent Variable Models, while NBEs are Neural-based Embeddings.

Experiments

Word sense induction

SemEval 2010 For the SemEval 2010 dataset, we compare models using two unsupervised metrics: V-measure (**V-M**) and paired F-score (**F-S**). V-M favors a high number of senses (e.g. assigning one cluster per instance), while F-S favors a small number of senses (e.g. all instances in one cluster) (Manandhar et al. 2010). In order to get a common ground for comparison, we do a geometric average AVG of both metrics, following (Wang et al. 2015). Finally, we also report the absolute difference between the actual (3.85) and induced number of senses as $\delta(\#S)$.

We compare with seven other models: a) LDA on co-occurrence graphs (**LDA**) and b) spectral clustering on co-occurrence graphs (**Spectral**) as reported in (Goyal and Hovy 2014), c) Hidden Concept (**HC**), d) HC using Zipf's law (**HC+Zipf**), and e) Bayesian nonparametric version of HC (**BNP-HC**) as reported in (Chang, Pei, and Chen 2014), f) CRP-based sense embeddings with positive PMI vectors as pre-trained vectors (**CRP-PPMI**), and g) Multi-Sense Skip-gram Model (**SE-WSI-fix**) as reported in (Song 2016).

Results are shown in Table 2a, where GAS outperforms other competing models on AVG. Among the GAS models, the GAS^{-wp} and GAS^{-sw} version perform the worst, emphasizing the necessity of the target-neighbor pairs and the switch variable. The overestimated GAS$^{s=100}$ performs better than previously proposed models, proving the robustness of our model on the different word sense granularities. On the $\delta(\#S)$ metric, the untuned GAS and GAS$^{s=5}$ perform the best. The V-M metric needs to be interpreted carefully, because it can easily be maximized by separating all instances into different sense clusters and thus overestimating the actual number of senses $\#S$ and decreasing the F-S metric. The model BNP-HC is an example of such: Though its V-M metric is the highest, it scores the lowest on the F-S metric and greatly overestimates $\#S$, thus having a very high $\delta(\#S)$. The goal is thus a good balance of V-M and F-S (i.e. highest AVG), and a close estimation of $\#S$ (i.e. lowest $\delta(\#S)$, which is successfully achieved by our models.

SemEval 2013 Two metrics are used for the SemEval 2013 dataset: fuzzy B-cubed (**F-BC**) and fuzzy normalized mutual information (**F-NMI**). F-BC gives preference to labelling all instances with the same sense, while F-NMI gives preference to labelling all instances with distinct senses. Therefore, computing the AVG of both metrics is also necessary in this experiment, for ease of comparison, as also suggested in (Wang et al. 2015).

We use seven baselines: a) lexical substitution method (**AI-KU**) and b) nonparametric HDP model (**Unimelb**) as reported in (Jurgens and Klapaftis 2013), c) Sense-Topic Model **STM**, d) STM with word2vec weights (**STM+w2v**) as reported in (Wang et al. 2015), e) Word Graph embeddings (**WG**), f) Distributional Inclusion Vector Embedding (**DIVE**) as reported in (Chang et al. 2018), and g) Multi Context Continuous model **MCC** as reported in (Komninos and Manandhar 2016).

Results are shown in Table 2b. Among the models, all versions of GAS perform better than other models on AVG. The untuned GAS and GAS$^{s=7}$ especially garner noticeable increase of 6.1% on fuzzy B-cubed metric from MCC, the previous best model. We also notice a big 6.0% decrease on the fuzzy B-cubed of GAS when the target-neighbor pair context is removed. This means that introducing the target-neighbor pair is crucial to the improvement of the model. Finally, the overestimated GAS model performs as well as the other GAS models, even outperforming all previous models on AVG, which proves the effectiveness of GAS even when s is set to a large value.

For completeness, we also report STM with additional contexts, STM+actual and STM+ukWac (Wang et al. 2015), where they used the actual additional contexts from the original data and semantically similar contexts from ukWac, respectively, as additional global context. With the performance gain we achieved, GAS without additional context can perform comparably to models with additional contexts: Our model greatly outperforms these models on the F-BC metric by at least 2%. Also, considering that both GAS and STM are LDA-based models, the same data enhancements

Sense	Word distribution	#Docs
1	hotel tour tourist summer flight	22
2	month ticket available performance	3
3	guest office stateroom class suite	3
*	advance overseas line popular japan	0
*	email day buy unable tour	0
*	sort basic tour time	0

Table 3: Six of the 15 senses of the target verb *book* using GAS with $S = 15$. The word lists shown are preprocessed to remove stopwords and the target word. The first three senses are senses which are assigned at least once to an instance document. The last three are *garbage senses*.

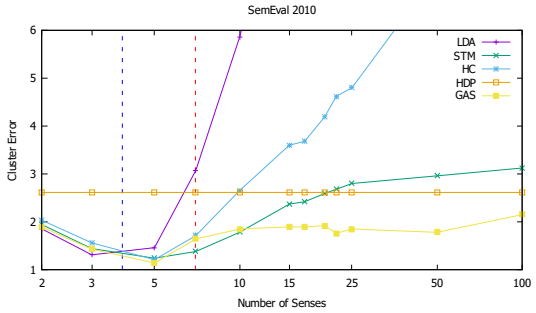
can straightforwardly be applied when the needs arise. We similarly apply the actual additional contexts to GAS and find that we achieve state-of-the-art performance on AVG.

Sense granularity problem

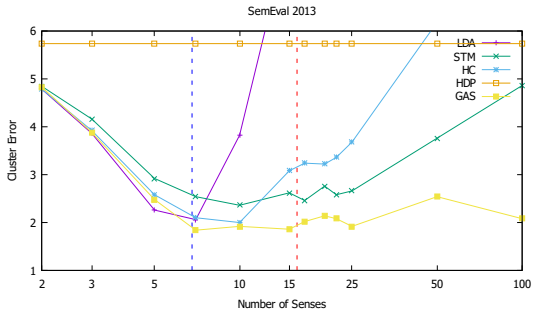
The main weakness of LDA when used on WSI task is the sense granularity problem. Recent models such as HC (Chang, Pei, and Chen 2014) and STM (Wang et al. 2015) mitigated this problem by tuning the number of senses hyperparameter S to minimize the error. However, such tuning, often empirically set to a small number such as $S = 3$ (Wang et al. 2015), fails to infer varying number of senses of words, especially for words with a higher number of senses. Nonparametric models such as HDP and BNP-HC (Lau, Cook, and Baldwin 2013; Chang, Pei, and Chen 2014) claim to automatically induce different S for each word. However, as shown in the results in Table 2, the estimated S is far from the actual number of senses and both models are ineffective.

On the other hand, Table 2 also shows that GAS is effective even when S is overestimated. We explain why through an example result shown in Table 3, where the target word is the verb *book*, the actual number of senses is three, and S is set to 15. First, we see that there are senses which are not assigned to any instance document, signified by $*$, which we call **garbage senses**. We notice that effectively representing a new latent variable for sense as a distribution over topics forces the model to throw garbage senses. Second, while it is easy to distinguish the third sense (i.e., *book: register in a booker*) to the two other senses, the first and second senses both refer to planning or arranging for an event in advance. Incorporating the target-neighbor pairs helps the model differentiates both into **fine-grained senses** *book: arrange for and reserve in advance* and *book: engage for a performance*.

We compare the competing models quantitatively on how they correctly detect the actual number of sense clusters using **cluster error**, which is the mean absolute error between the detected number and the actual number of sense clusters. We compare the cluster errors of LDA (Blei, Ng, and Jordan 2003), STM (Wang et al. 2015), HC (Chang, Pei, and Chen 2014), and a nonparametric model HDP (Teh et al. 2004), with GAS. We report the results in Figure 4. Results show that the cluster error of LDA increases sharply as the number of senses exceeds the actual mean number of senses. HC and STM also throw garbage senses since they



(a) SemEval 2010 dataset



(b) SemEval 2013 dataset

Figure 4: Cluster error of models with different number of senses S. The vertical dashed lines correspond to the mean and the max of the actual number of senses. The x-axes are log-scaled.

also introduce in some way a new sense variable, however the cluster errors of both models still increase when S is set beyond the maximum number of senses. We argue that this is because first, the sense representation is not optimal as they assume strict local/global context assumption, and second and most importantly, the models do not produce fine-grained senses. GAS does both garbage sense throwing and fine-grained sense induction, which helps in the detection of the actual word granularity. Finally, the cluster error of GAS is always better than that of HDP. This shows that GAS, despite being a parametric model, automatically detects the number of sense clusters without parameter tuning and is more accurate than the automatic detection of nonparametric models.

Unsupervised author name disambiguation

Unsupervised author name disambiguation (UAND) is a task very similar to the WSI task, where ambiguous author names are the target words. However, one additional challenge of UAND is that there can be as many as 100 authors with the same name, whereas words can have at most 20 different senses, at least in our datasets, as shown in the dataset statistics in Table 4. Moreover, the standard deviations of the author name disambiguation datasets are also higher, which means that there is more variation on the number of senses per target author name. Thus, in this task, the sense granularity problem is more difficult and needs to be addressed

Dataset	Min	Max	Mean	StdDev
SemEval 2010	2	16	7.68	3.35
SemEval 2013	2	7	3.85	1.40
PubMed	1	28	10.41	7.68
Arnet	1	112	14.18	18.02

Table 4: Statistics of the number of senses of target words/names in the datasets used in the paper.

properly.

Current state-of-the-art models use non-text features such as publication venue and citations (Tang et al. 2012). We argue that text features also provide informative clues to disambiguate author names. In this experiment, we make use of text features such as the title and abstract of research papers as data instance of the task. In addition, we also include in our dataset author names and the publication venue as pseudo-words. In this way, we can reformulate the UAND task as a WSI task, and exploit text features not used in current techniques.

Experimental setup We use two publicly available datasets for the UAND task: Arnet[4] and PubMed[5]. The Arnet dataset contains 100 ambiguous author names and a total of 7528 papers as data instance. Each instance includes the title, author list, and publication venue of a research paper authored by the given author name. In addition, we also manually extract the abstracts of the research papers for additional context. The PubMed dataset contains 37 author names with a total of 2875 research papers as instances. It includes the PubMed ID of the papers authored by the given author name. We extract the title, author list, publication venue, and abstract of each PubMed ID from the PubMed website.

We use LDA (Blei, Ng, and Jordan 2003), HC (Chang, Pei, and Chen 2014) and STM (Wang et al. 2015) as baselines. We do not compare with non-text feature-based models (Tang et al. 2012; Cen et al. 2013) because our goal is to compare sense topic models on a task where the sense granularities are more varied. For STM and GAS, the title, publication venue and the author names are used as local contexts while the abstract is used as the global context. This decision is based on conclusions from previous works (Tang et al. 2012) that the title, publication venue, and the author names are more informative than the abstract when disambiguating author names. We use the same parameters as used above, and we set S to 5, 25, 50, and 100 to compare the performances of the models as the number of senses increases.

Results For evaluation, we use the pairwise F1 measure to compare the performance of competing models, following (Tang et al. 2012). Results are shown in Figure 5. GAS performs the best on almost all settings, except on the PubMed dataset and when $S = 5$, where it garners a comparable result with STM. However, in the case where S is set close to the maximum number of senses in the dataset (i.e. 28 in PubMed and 112 in Arnet), GAS performs the best among the models. LDA and HC perform badly on all settings and greatly decrease their performances when S becomes high. STM also shows decrease in performance on the PubMed dataset when $S = 100$. This is because the PubMed dataset has a lower maximum number of senses, and STM is sensitive in the setting of S, and thus hurts the robustness of the model to different sense granularities.

[4] https://aminer.org/disambiguation

[5] https://github.com/Yonsei-TSMM/author_name_disambiguation

Model	$S = 5$	$S = 25$	$S = 50$	$S = 100$
LDA	31.5	13.4	9.8	8.2
HC	46.3	46.3	44.4	41.7
STM	52.8	55.0	55.5	55.0
GAS	56.2	56.4	57.9	58.8

(a) Arnet Dataset

Model	$S = 5$	$S = 25$	$S = 50$	$S = 100$
LDA	41.4	13.3	8.9	9.0
HC	42.5	44.1	41.6	41.3
STM	44.9	44.4	44.9	41.9
GAS	44.4	45.5	46.6	46.5

(b) PubMed Dataset

Table 5: Paired F1 measures of competing models with different number of senses S on UAND datasets.

Conclusion

We proposed a solution to answer the sense granularity problem, one of the major challenges of the WSI task. We introduced GAS, a latent variable model that not only throws away garbage senses, but also induces fine-grained senses. We showed that GAS greatly outperforms the current state-of-the-art models in both SemEval 2010 and 2013 WSI datasets. We also show experiments on how GAS is able to overcome sense granularity problem, a well-known flaw of latent variable models on. We further applied our model to UAND task, a similar task but with more varying number of senses, and showed that GAS performs the best among latent variable models, proving its robustness to different sense granularities.

References

Almuhareb, A.; Poesio, M.; et al. 2006. Msda: Wordsense discrimination using context vectors and attributes. In *ECAI*, 543–547.

Baskaya, O.; Sert, E.; Cirik, V.; and Yuret, D. 2013. Aiku: Using substitute vectors and co-occurrence modeling for word sense induction and disambiguation. *SemEval* 300–306.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Brody, S., and Lapata, M. 2009. Bayesian word sense induction. In *EACL*, 103–111. Association for Computational Linguistics.

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

Cen, L.; Dragut, E. C.; Si, L.; and Ouzzani, M. 2013. Author disambiguation by hierarchical agglomerative clustering with adaptive stopping criterion. In *SIGIR*, 741–744. ACM.

Chang, H.-S.; Agrawal, A.; Ganesh, A.; Desai, A.; Mathur, V.; Hough, A.; and McCallum, A. 2018. Efficient graph-based word sense induction by distributional inclusion vector embeddings. *arXiv preprint arXiv:1804.03257*.

Chang, B.; Pei, W.; and Chen, M. 2014. Inducing word sense with automatically learned hidden concepts. In *COLING*, 355–364.

Chemudugunta, C.; Smyth, P.; and Steyvers, M. 2006. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS*, volume 19, 241–248.

Goyal, K., and Hovy, E. H. 2014. Unsupervised word sense induction using distributional statistics. In *COLING*, 1302–1310.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences* 101(suppl 1):5228–5235.

Heckerman, D.; Chickering, D. M.; Meek, C.; Rounthwaite, R.; and Kadie, C. 2000. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research* 1(Oct):49–75.

Jurgens, D., and Klapaftis, I. 2013. Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In **SEM*, volume 2, 290–299.

Komninos, A., and Manandhar, S. 2016. Structured generative models of continuous features for word sense induction. *COLING* 11.

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

Lau, J. H.; Cook, P.; and Baldwin, T. 2013. unimelb: Topic modelling-based word sense induction for web snippet clustering. In *SemEval*, 217–221.

Manandhar, S.; Klapaftis, I. P.; Dligach, D.; and Pradhan, S. S. 2010. Semeval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th international workshop on semantic evaluation*, 63–68. Association for Computational Linguistics.

Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; and McClosky, D. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55–60.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, 3111–3119.

Pelevina, M.; Arefiev, N.; Biemann, C.; and Panchenko, A. 2016. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, 174–183.

Shu, L.; Long, B.; and Meng, W. 2009. A latent topic model for complete entity resolution. In *ICDE*, 880–891. IEEE.

Song, Y.; Huang, J.; Councill, I. G.; Li, J.; and Giles, C. L. 2007. Efficient topic-based unsupervised name disambiguation. In *JCDL*, 342–351. ACM.

Song, L. 2016. Word embeddings, sense embeddings and their application to word sense induction. *The University of Rochester, April*.

Stevenson, M., and Wilks, Y. 2003. Word sense disambiguation. *The Oxford Handbook of Comp. Linguistics* 249–265.

Tang, J.; Fong, A. C.; Wang, B.; and Zhang, J. 2012. A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering* 24(6):975–987.

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2004. Sharing clusters among related groups: Hierarchical dirichlet processes. In *NIPS*, 1385–1392.

Tsvetkov, Y.; Schneider, N.; Hovy, D.; Bhatia, A.; Faruqui, M.; and Dyer, C. 2014. Augmenting english adjective senses with supersenses. In *LREC*.

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

Yao, X., and Van Durme, B. 2011. Nonparametric bayesian word sense induction. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, 10–14. Association for Computational Linguistics.

Yao, X.; Van Durme, B.; and Callison-Burch, C. 2012. Expectations of word sense in parallel corpora. In *NAACL-HLT*, 621–625. Association for Computational Linguistics.