Inducing Word Sense with Automatically Learned Hidden Concepts

Baobao Chang **Wenzhe Pei** **Miaohong Chen**
Key Laboratory of Computational Linguistics, Ministry of Education
School of Electronics Engineering and Computer Science, Peking University
Beijing, P.R.China, 100871
{chbb,peiwenzhe,miaohong-chen}@pku.edu.cn

Abstract

Word Sense Induction (WSI) aims to automatically induce meanings of a polysemous word from unlabeled corpora. In this paper, we first propose a novel Bayesian parametric model to WSI. Unlike previous work, our research introduces a layer of hidden concepts and view senses as mixtures of concepts. We believe that concepts generalize the contexts, allowing the model to measure the sense similarity at a more general level. The Zipf's law of meaning is used as a way of pre-setting the sense number for the parametric model. We further extend the parametric model to non-parametric model which not only simplifies the problem of model selection but also brings improved performance. We test our model on the benchmark datasets released by Semeval-2010 and Semeval-2007. The test results show that our model outperforms state-of-the-art systems.

1 Introduction

Word Sense Induction (WSI) aims to automatically induce meanings of a polysemous word from unlabeled corpora. It discriminates among meanings of a word by identifying clusters of similar contexts. Unlike the task of Word Sense Disambiguation (WSD), which classifies polysemous words according to a pre-existing and usually hand-crafted inventory of senses, WSI makes it attractive to researchers by eliminating dependence on a particular sense inventory and learning word meaning distinction directly based on the contexts as observed in corpora.

Almost all WSI work relies on the distributional hypothesis, which states that words occurring in similar contexts will have similar meanings. To effectively discriminate among contexts, proper representation of contexts would be a key issue. Basically, context can be represented as a vector of words co-occurring with the target word within a fixed context window. The similarity between two contexts of the target word can then be measured by the geometrical distance between the corresponding vectors. To ease the sparse problem and capture more semantic content, some kinds of generalizations or abstractions are needed. For example, a context of *bank* including *money* may not share similarity with that including *cash* measured at word level. However, given the conceptual relationship between *money* and *cash*, the two contexts actually share high similarity.

One straightforward way of introducing conceptualization is to assign semantic code to context words, where semantic codes could be derived from WordNet or other resources like thesauruses. However, two problems remain to be tackled. The first one concerns ambiguities of context words. Context words may have multiple semantic codes and thus word sense disambiguation to context words or other extra cost is needed. The second one concerns the nature of WSI task. WSI actually is target-word-specific, which means the conceptualization should be done specifically to different target words. A general purpose conceptualization defined by a thesaurus may not well meet this requirement and may not be equally successful in discriminating contexts of different target words.

To address these problems, we first propose a parametric Bayesian model which jointly finds conceptual representations of context words and the sense of the target word. We do this by introducing a layer of target-specific conceptual representation between the target sense layer and the context words layer

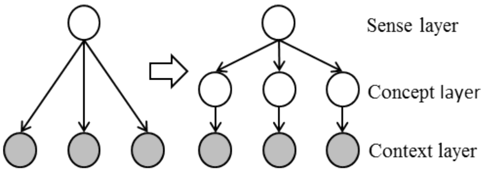

Figure 1: Architecture of our model

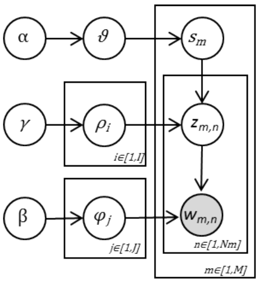

Figure 2: Graphical notation of the Basic Model

through a Bayesian framework as illustrated in Figure 1. From the generative perspective, the sense of the target word is first sampled. Then the sense generates different conceptual configurations which in turn generate different contexts. With a deeper architecture, our model makes it possible to induce word senses at a more abstract level, i.e. the concept level, which is not only less sparse but also more semantically oriented. Both the senses of the target word and the latent concepts are inferred automatically and unsupervisedly with inference procedure given enough contexts involving a target word. The latent concepts inferred with the model share similarities with those defined in thesauruses, as both of them cluster semantically related words. However, since the latent concepts are inferred with regard to individual target words, they are target-word-specific and thus fit the WSI task better than general purpose concepts defined in thesauruses. Context words may still correspond to multiple latent concepts. However, the disambiguation is implicitly done in the process of the word sense induction.

Setting the number of senses that the algorithm should arrive at is another problem frequently exercising the minds of WSI people. Instead of trying different sense numbers on a word-by-word basis, we propose to use Zipf's law of meaning (Zipf, 1945) to guide the selection of the sense numbers in this paper. With the law of meaning, sense numbers could be set on an all-word basis, rather than on a word-by-word basis. This is not only simple but also efficient, especially in the case where there are a large number of target words to be concerned.

We further extend the parametric model into a non-parametric model, as it allows adaptation of model complexity to data. By extending our model to non-parametric model, the need to preset the numbers of senses and latent concepts are totally removed and, moreover, the model performance is also improved.

We evaluate our model on the commonly used benchmark datasets released by both Semeval-2010 (Manandhar et al., 2010) and Semeval-2007 (Agirre and Soroa, 2007). The test results show that our models perform much better than the state-of-the-art systems.

2 The parametric model

2.1 Basic Model

The main point of our work is that different senses are signaled by contexts with different concept configurations, where different concepts are formally defined as different distributions over context words. Formally, we denote by $P(s)$ the global multinomial distribution over senses of an ambiguous word and by $P(w|z)$ the multinomial distributions over context words w given concept z. Context words are generated by a mixture of different concepts whose mixture proportion is defined by $P(z|s)$, such that:

$$P(w_i) = \sum_j P(s = j) \sum_k P(z_i = k|s = j) P(w_i|z_i = k)$$

Following the model, each context word w_i surrounding the target word is generated as follows: First, a sense s is sampled from $P(s)$ for the target word. Then for each context word position i, a concept z_i is sampled according to mixture proportion $P(z|s)$ and w_i is finally sampled from $P(w|z)$.

Figure 2 shows the model with the graphical notation, where M is the number of instances of contexts regarding to a concerned target word and N_m is the number of word tokens in context m. s_m is the sense label for target word in context m. $w_{m,n}$ is the n-th context word in context m. $z_{m,n}$ is the concept label associated with $w_{m,n}$. I is the total number of senses to be induced. J is the total number

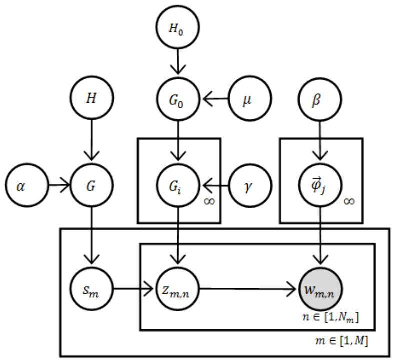

Figure 3: Graphical notation of the non-parametric WSI model

of concepts. $\vec{\theta}$ is the notational shorthand for the sense distribution $P(s)$, $\vec{\rho}_i$ is the shorthand for the i-th sense-concept distribution $P(z|s = i)$, and $\vec{\varphi}_j$ is the j-th concept-word distribution $P(w|z = j)$. Following conventional Bayesian practice, $\vec{\theta}$, $\vec{\rho}_i$ and $\vec{\varphi}_j$ are assumed to be drawn from Dirichlet priors with symmetric parameter α, γ, β respectively. The observed variable is represented with shaded node and hidden variable with unshaded node.

2.2 Zipf's law of meaning

Most of the WSI work requires that the number of senses to be induced be specified ahead of time. One straightforward way to deal with this problem is to repeatedly try different numbers of senses on a development set and select the best performed number. However, this should be done in principle on a word-by-word basis, and thus could be time-consuming and prohibitive when there are lots of target words to be concerned. A more systematic way of setting sense numbers in Bayesian models is extending the parametric model into a non-parametric model, which will be described in detail in section 3.

To work with our parametric model, we propose in this paper that an empirical law, Zipf's law of meaning (Zipf, 1945), could be used to guide the sense number selection. Zipf's law of meaning states that the number of sense of a word is proportional to its frequency as shown in the following equation:

$$I = K * f^b \tag{1}$$

where I is the number of word senses and f is the frequency of the word. K is the coefficient of proportionality which is unknown and b is about 0.404 according to an experimental study done by Edmonds (2006).

Certainly, Zipf's law of meaning is not as strict as a rigorous mathematical law. However, it sketches the distribution of the sense numbers with word frequencies of all words and allows us to estimate the sense numbers on an all-word basis by selecting appropriate coefficient K. This is not only simple but also efficient, especially in the case that there are a large number of target words to be concerned.

3 Non-parametric Model

A limitation of the parametric model is that the sense number I of the target word and the number J of latent concepts need to be fixed beforehand. Bayesian non-parametric (BNP) models offer elegant approach to the problem of model selection and adaption. Rather than comparing models that vary in complexity, the BNP approach is to fit a single model that can adapt its complexity to the data. Unlike the parametric approach, BNP approach assumes an infinite number of clusters, among which only a few are active given the training data. Our basic model can be naturally extended into a BNP model as shown in Figure 3. Instead of assuming a finite number of senses, we place a nonparametric, Dirichlet process (DP) prior on the sense distribution as follows:

$$\begin{aligned} G &\sim DP(\alpha, H) \\ s_m &\sim G, m = 1, 2, \ldots, M \end{aligned}$$

where α is the concentration parameter and H is the base measure of the Dirichlet process.

For each sense s_i of the target words, we place a Hierarchical Dirichlet process (HDP) prior on the mixture proportion to latent concepts shown as follows:

$$\begin{aligned} G_0 &\sim DP(\mu, H_0) \\ G_i &\sim DP(\gamma, G_0), i = 1, 2, \ldots \\ z_{m,n} &\sim G_i, n = 1, 2, \ldots, N_m \\ w_{m,n} &\sim \vec{\varphi}_{z_{m,n}} \end{aligned}$$

where μ and γ are concentration parameters to G_0 and G_i, H_0 is the base measure of G_0.

By using HDP priors, we make sure that the same set of concept-word distributions is shared across all senses and all contexts of a target word, since each random measure G_i inherits its set of concepts from the same G_0.

As in parametric model, $\vec{\varphi}_j$ is the j-th concept-word distribution $P(w|z = j)$, however, there are now an infinite number of such distributions. So is the number of senses. However, with a fixed number of contexts of the target word, only a finite number of senses and concepts are active and they could be inferred automatically by the inference procedure.

4 Model Inference

We use Gibbs sampling (Casella and George, 1992) for inference to both the parametric and nonparametric model. As a particular Markov Chain Monte Carlo (MCMC) method, Gibbs sampling is widely used for inference in various Bayesian models (Teh et al., 2006; Li and Li, 2013; Li and Cardie, 2014).

4.1 The Parametric Model

For the parametric model, we use collapsed Gibbs sampling, in which the sense distribution $\vec{\theta}$, sense-concept distribution $\vec{\rho}_i$ and concept-word distribution $\vec{\varphi}_j$ are integrated out. At each iteration, the sense label s_m of the target word in context m is sampled from conditional distribution $p(s_m|\vec{s}_{\neg m}, \vec{z}, \vec{w})$, and the concept label $z_{m,n}$ for the context word $w_{m,n}$ is sampled from conditional distribution $p(z_{m,n}|\vec{s}, \vec{z}_{\neg(m,n)}, \vec{w})$. Here $\vec{s}_{\neg m}$ refers to all current sense assignments other than s_m and $\vec{z}_{\neg(m,n)}$ refers to all current concept assignment other than $z_{m,n}$.

The conditional distribution $p(s_m|\vec{s}_{\neg m}, \vec{z}, \vec{w})$ and $p(z_{m,n}|\vec{s}, \vec{z}_{\neg(m,n)}, \vec{w})$ can be derived as shown in equation (2) and (3) respectively:

$$p(s_m = i|\vec{s}_{\neg m}, \vec{z}, \vec{w}; \alpha, \beta, \gamma) \propto (c_i^{\neg m} + \alpha) \cdot \frac{\prod_{j=1}^{J} \prod_{x=1}^{f_{m,j}} (c_{i,j}^{\neg m} + \gamma + x - 1)}{\prod_{x=1}^{f_{m,*}} (\sum_{j=1}^{J} c_{i,j}^{\neg m} + J * \gamma + x - 1)} \tag{2}$$

$$p(z_{m,n} = j|\vec{s}, \vec{z}_{\neg(m,n)}, \vec{w}; \alpha, \beta, \gamma) \propto (c_{s_m,j}^{\neg(m,n)} + \gamma) \cdot \frac{(c_{j,w_{m,n}}^{\neg(m,n)} + \beta)}{\sum_{t=1}^{V} c_{j,t}^{\neg(m,n)} + V * \beta} \tag{3}$$

Here, $c_i^{\neg m}$ is the number of instances with sense i. $c_{i,j}^{\neg m}$ is the number of concept j in instances with sense i. Both of them are counted without the m-th instance of the target word. $c_{s_m,j}^{\neg(m,n)}$ is defined in a similar way with $c_{i,j}^{\neg m}$ but without counting the word position (m, n). $c_{j,w_{m,n}}^{\neg(m,n)}$ is the number of times word $w_{m,n}$ is assigned to concept j without counting word position (m, n). $f_{m,j}$ is the number of concept j assigned to context words in instance m and $f_{m,*}$ is the total number of words in contexts of instance m. V stands for the size of the word dictionary, i.e. the number of different words in the data. x is an index which iterates from 1 to $f_{m,*}$.

$\vec{\theta}$, $\vec{\rho}_i$ and $\vec{\varphi}_j$ can be estimated in a similar way, we now only show as example the estimation of $\vec{\rho}_i$, parameters for sense-concept distributions. According to their definitions as multinomial distributions with Dirichlet prior, applying Bayes' rule yields:

$$p(\vec{\rho}_i|\vec{z}; \vec{\gamma}) = \frac{p(\vec{\rho}_i; \vec{\gamma}) * p(\vec{z}|\vec{\rho}_i; \vec{\gamma})}{Z_{\vec{\rho}_i}} = Dir(\vec{\rho}_i|\vec{c}_i + \vec{\gamma})$$

where $\vec{c_i}$ is the vector of concept counts for sense i. Using the expectation of the Dirichlet distribution, values of $\rho_{i,j}$ can be worked out as follows:

$$\rho_{i,j} = \frac{c_{i,j} + \gamma}{\sum_{k=1}^{J} c_{i,k} + J * \gamma}$$

Different read-outs of $\rho_{i,j}$ are then averaged to produce the final estimation.

4.2 The Non-parametric Model

Chinese restaurant process (CRP) and Chinese restaurant franchise (CRF) process (Teh et al., 2006) have been widely used as sampling scheme for DP and HDP respectively. As our non-parametric model involves both DP and HDP, we use both CRP and CRF based sampling for model inference.

In the CRP metaphor to DP, there is one Chinese restaurant with an infinite number of tables, each of which can seat an infinite number of customers. The first customer enters the restaurant and sits at the first table. The second customer enters and decides either to sit with the first customer or by herself at a new table. In general, the $n + 1$st customer either joins an already occupied table k with probability proportional to the number n_k of customers already sitting there, or sits at a new table with probability proportional to α. As in our model, when we sample the sense s_m for each context, we assume that tables correspond to senses of target words and customers correspond to whole contexts in which the target word occurs.

In the CRF metaphor to HDP, there are multiple Chinese restaurants, and each one has infinitely many tables. On each table the restaurant serves one of infinitely many dishes that other restaurants may serve as well. At each table of each restaurant one dish is ordered from the menu by the first customer who sits there, and it is shared among all customers who sit at that table. The menu is shared by all the restaurants. To be specific to our model, when we sample the concept $z_{m,n}$ for each context word, we assume each sense s_m of the target word corresponds to a restaurant and each word $w_{m,n}$ corresponds to a customer while concept $z_{m,n}$ corresponds to the dishes served to the customer by the restaurant. Neither the number of restaurant nor the number of dishes is finite in our model.

For model inference, we first sample s_m using CRP-based sampling and then we sample $z_{m,n}$ for each s_m using CRF-based sampling. The sampling of s_m and $z_{m,n}$ are done alternately, but not independently. The sampling of s_m is conditional on the current value of $z_{m,n}$ and vice versa, conforming to the scheme of Gibbs Sampling.

The equation for sampling s_m is derived as in equation (4):

$$p(s_m = i|\vec{s}_{\neg m}, \vec{z}, \vec{w}) \propto \begin{cases} c_i^{\neg m} \cdot p(\vec{z}_m|\vec{z}_{\neg m}, s_m = i) & \textbf{if} \quad i = old \\ \alpha \cdot p(\vec{z}_m|\vec{z}_{\neg m}, s_m = i^{new}) & \textbf{else} \end{cases}$$

$$where$$

$$p(\vec{z}_m|\vec{z}_{\neg m}, s_m = i) = \frac{\prod_{j=1}^{J} \prod_{x=1}^{f_{m,j}} (c_{i,j}^{\neg m} + \gamma * \frac{c_{t,j}^{\neg m}}{c_{t,*}^{\neg m} + \mu} + x - 1)}{\prod_{x=1}^{f_{m,*}} (\sum_{j=1}^{J} c_{i,j}^{\neg m} + \gamma + x - 1)} \tag{4}$$

Here $p(\vec{z}_m|\vec{z}_{\neg m}, s_m = i)$ is estimated block-wise for context m according to the CRF metaphor. $c_i^{\neg m}$ and $c_{i,j}^{\neg m}$ are defined in the same way as that in equation (2). $c_{t,j}^{\neg m}$ is the number of tables with dish j in all restaurants but m and $c_{t,*}^{\neg m}$ means the number of tables in all restaurants but m. x is an index which iterates from 1 to $f_{m,*}$.

Sampling $z_{m,n}$ needs more steps than sampling s_m as we need to record the table assignment for each dish (concept). For each dish $z_{m,n}$ of a customer $w_{m,n}$, we first sample the table at which the customer sits according to the following equations:

$$p(t_{m,n} = t|\vec{t}_{\neg(m,n)}, \vec{z}_{\neg(m,n)}, w_{m,n}, s_m = i) \propto \begin{cases} c_{i,t}^{\neg(m,n)} \cdot p_j^{\neg(m,n)}(w_{m,n}) & \textbf{if} \quad t = old \\ \gamma \cdot p(w_{m,n}|\vec{t}_{\neg(m,n)}, t_{m,n} = t, \vec{z}_{\neg(m,n)}, w_{m,n}) & \textbf{else} \end{cases}$$

$$where$$

$$p_j^{\neg(m,n)}(w_{m,n}) = p(w_{m,n}|z_{m,n} = j, \vec{w}_{\neg(m,n)}) = \frac{c_{j,w_{m,n}}^{\neg(m,n)} + \beta}{\sum_{t=1}^{V} c_{j,t}^{\neg(m,n)} + V\beta}$$

	Basic Model	BNP
α	1.0	0.2
β	0.05	0.01
γ	0.05	0.2
μ	N/A	0.001
K	0.27	N/A
Concept number	20	N/A
Context window	± 5 words	± 9 words

Table 1: Hyperparamters of our models

Here $c_{i,t}^{\neg(m,n)}$ is the number of customers on table t in restaurant i and $c_{j,w_{m,n}}^{\neg(m,n)}$ has the same meaning as in equation (3). If the sampled table t is previously occupied, then $z_{m,n}$ is set to the dish j assigned to t according to the CRF metaphor. If the sampled table t is new, the probability $p(w_{m,n}|\vec{t}_{\neg(m,n)}, t_{m,n} = t, \vec{z}_{\neg(m,n)}, w_{m,n})$ is calculated using equation (5), which is the sum of the probability of all previously ordered dishes and the newly ordered dish.

$$p(w_{m,n}|\vec{t}_{\neg(m,n)}, t_{m,n} = t, \vec{z}_{\neg(m,n)}, w_{m,n}) = \sum_{j=1}^{J} \frac{c_{t,j}^{\neg(m,n)}}{c_{t,*}^{\neg(m,n)} + \mu} \cdot p_j^{\neg(m,n)}(w_{m,n}) + \frac{\mu}{c_{t,*}^{\neg(m,n)} + \mu} \cdot p_{j^{new}}^{\neg(m,n)} \tag{5}$$

Because a new table is added, we then sample a new dish for this table according to equation (6).

$$p(z_{m,n} = j|\vec{t}, \vec{z}_{\neg(m,n)}) \propto \begin{cases} c_{t,j}^{\neg(m,n)} \cdot p_j^{\neg(m,n)}(w_{m,n}) & \textbf{if} \quad j = old \\ \mu \cdot p_{j^{new}}^{\neg(m,n)}(w_{m,n}) & \textbf{if} \quad j = new \end{cases} \tag{6}$$

After the dish j is sampled, it is assigned to the new table and the number of table serving dish j is added.

Parameters $\vec{\theta}$, $\vec{\rho}_i$ and $\vec{\varphi}_j$ can be estimated in the same way as described in section 4.1.

5 Experiment

5.1 Experiment Setup

Data Our primary WSI evaluation is based on the standard dataset in Semeval-2010 Word sense induction & Disambiguation task (Manandhar et al., 2010). The target word dataset consists of 100 words, 50 nouns and 50 verbs. There are a total number of 879,807 sentences in training set and 8,915 sentences in testing set. The average number of word senses in the data is 3.79.

Model Selection The trail data of Semeval-2010 WSI task is used as development set for parameter tuning, which consists of training and test portions of 4 verbs. The 4 verbs are different words than the 100 target words in the training data. There are only about 138 instances on average for each target word in the training part of the trial data. To make a development set of more reasonable size, the trial data are supplemented with 6K instances of the 4 verbs extracted from the British National Corpus (BNC)[1] corpus. As we use the Zipf's law of meaning to guide the selection of number of senses, BNC was also used to count word frequencies.

The final hyper-parameters are set as in Table 1. In all the following experiments, Gibbs sampler is run for 2000 iterations with burn-in period of 500 iterations. Every 10th sample is read out for parameter estimating after the burn-in period to avoid autocorrelation. Due to the randomized property of Gibbs sampler, all results in the next sections are averaged over 5 runs. The average running time for each target word is about 7 minutes on a computer equipped with an Intel Core i5 processor working at 3.1GHz and 8GB RAM.

Pre-Processing For each instance of the target word in training data and testing data, all words are lemmatized and stop words like '*of*', '*the*', '*a*' which are irrelevant to word sense distinction are filtered. Words occurring less than twice are removed.

Evaluation method Semeval-2010 WSI task presents two evaluation schemes which are supervised evaluation and unsupervised evaluation. In supervised evaluation, the gold standard dataset is split into

[1] www.natcorp.ox.ac.uk/

Model	Supervised Evaluation		Unsupervised Evaluation		Averaged #s
	80-20 split	60-40 split	V-Measure	Paired-Fscore	
Basic Model	64.12	63.68	11.52	**44.42**	5
Basic Model + Zipf	66.4	65.25	15.2	35.12	7.66
BNP	**69.3**	**68.9**	**21.4**	23.1	15.62

Table 2: Test results with different configurations.

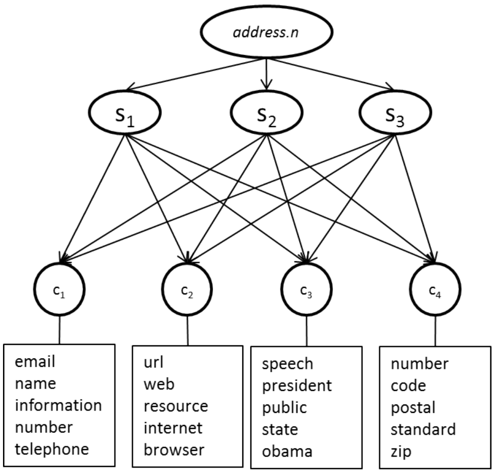

Figure 4: Examples of concepts induced with the BNP model specific to the target word *address.n* (with c_i denoting concept)

a mapping and an evaluation parts. The first part is used to map the automatically induced senses to gold standard senses. The mapping is then used to calculate the system's F-Score on the second part. According to the size of mapping data and evaluation data, the evaluation results are measured on two different splits which are 80-20 splits and 60-40 splits. 80-20 splits means that 80% of the test data are used for mapping and 20% are used for evaluation. In unsupervised evaluation, the system outputs are compared by using metrics V-Measure (Rosenberg and Hirschberg, 2007) and Paired F-Score (Artiles et al., 2009).

5.2 Experiment Results

Table 2 lists all experiment results. The Basic Model stands for the parametric model with fixed number of senses for all target words. The number of senses is set to 5 which gives the best performance on development set. Basic Model + Zipf is the model with the number of sense estimated by Zipf's law of meaning. BNP stands for our non-parametric model. As we can see, compared with the Basic Model with fixed sense number, the model using Zipf's law of meaning achieves improved performance. This means Zipf's law of meaning has positive effect in setting the sense number of the WSI task. BNP achieves the best performance on both supervised evaluation and V-measure evaluation. In terms of Paired F-score, however, the Basic Model gets the best results while BNP performs worst. This is consistent with what claimed by Manandhar et al. (2010), that Paired F-score tends to penalize the model with higher number of clusters.

As stated before, our models not only perform word sense induction but also group the context words into concepts. Figure 4 shows 4 of the concepts induced by BNP with regard to the target word *address.n*. Senses of *address.n* are defined as the mixture of concepts and concepts are defined as distributions over context words. We only list the top five words with the highest probabilities under each concept. As shown in Table 2, the non-parametric model induces much finer granularity of senses than the gold standard, it makes distinction among *email address*, *web address*, and even *ip address*. A possible solution is to further measure the closeness of senses based on the sense representations induced and merge similar senses to produce coarser granularity of senses.

Model	F-score(%)
BNP+position	**69.7**
BNP	69.3
Basic Model + Zipf	66.4
Basic Model	64.1
HDP	65.8
HDP+position (Lau et al., 2012)	68
distNB (Choe and Charniak, 2013)	65.4
UoY (Korkontzelos and Manandhar, 2010)	62.4

Model	F-score(%)
BNP+position	**88.0**
BNP	86.1
HDP (Yao and Van Durme, 2011)	85.7
HDP+position (Lau et al., 2012)	86.9
Feature-LDA (Brody and Lapata, 2009)	85.5
1-layer-LDA (Brody and Lapata, 2009)	84.6
HRG (Klapaftis and Manandhar, 2010)	87.6
I2R (Niu et al., 2007)	86.8

Table 3: Comparison with state-of-the-arts on Semeval-2010 data (left) and Semeval-2007 data (right)

5.3 Comparison with previous work

Much previous work (Brody and Lapata, 2009; Klapaftis and Manandhar, 2010; Yao and Van Durme, 2011) tested their models only on Semeval-2007 dataset (Agirre and Soroa, 2007) which consists of roughly 27K instances of 65 target verbs and 35 target nouns, coming from the Wall Street Journal corpus (WSJ) (Agirre and Soroa, 2007). For a complete comparison, we also test our model on the Semeval-2007 dataset. Since training data was not provided as part of the original Semeval-2007 dataset, we follow the approach of previous work (Brody and Lapata, 2009; Yao and Van Durme, 2011; Lau et al., 2012) to construct training data for each target word by extracting instances from the BNC corpus. Following paractices as much previous work (Brody and Lapata, 2009; Yao and Van Durme, 2011; Lau et al., 2012) did, we compare with previous work with supervised F-score on 80-20 data split in Semeval-2010 and noun data in Semeval-2007.

Table 3 (left) compares our models against the state-of-the-art systems tested on 80-20 data split in Semeval-2010. HDP+position (Lau et al., 2012) improved the HDP model (Yao and Van Durme, 2011) by including a position feature. distNB (Choe and Charniak, 2013) extends the naive Bayes model by reweighting the conditional probability of a context word given the sense by its distance to the target word. UoY (Korkontzelos and Manandhar, 2010) is the best performing system in Semeval-2010 competition which used a graph-based model. We re-implemented and tested the HDP model on the Semeval-2010 dataset since Yao and Van Durme (2011) and Lau et al. (2012) did not report their HDP results on this dataset.

Different with normal practice in WSI work, there is no feature engineering in our model. However, our BNP model outperformed all the systems on supervised evaluation. Even the Basic Model outperformed the best performing Semeval-2010 system. Especially, our BNP model performs much better than the HDP model. Both Lau et al. (2012) and Choe and Charniak (2013) show benefit of using positional information. Since our model does not exclude further feature engineering, we also introduce a position feature[2] into our non-parametric model (**BNP+position**) as in Lau et al. (2012). This contributes to a further 0.4% rise in performance.

Table 3 (right) compares our models with previous work on the nouns dataset in Semeval-2007. We divides systems being compared into two groups. The first group model the WSI task with Bayesian framework, while the second group uses models other than Bayesian model. Feature-LDA is the LDA-based model proposed by Brody and Lapata (2009) which incorporates a large number of features into the model. The 1-layer-LDA is their model with only bag-of-words features. HRG is a hierarchical random graph model. I2R is the best performing system in Semeval-2007. As shown in Table 3 (right), our BNP model with position feature (**BNP+position**) outperforms all systems. If we restrict our attention to the first group in which all models are Bayesian model, our BNP model without feature engineering outperforms the HDP model which is also non-parametric model without feature engineering.

6 Related Work

A large body of previous work is devoted to the task of Word Sense Induction. Almost all work relies on the distributional hypothesis, which states that words occurring in similar contexts will have similar meanings. Different work exploits distributional information in different forms, including context clustering models (Schütze, 1998; Niu et al., 2007; Pedersen, 2010; Elshamy et al., 2010; Kern et al., 2010), graph-based models (Korkontzelos and Manandhar, 2010; Klapaftis and Manandhar, 2010) and Bayesian

[2] Formally, the position feature is the context words with its relative position to the target word.

models. For Bayesian models, Brody and Lapata (2009) firstly introduced a Bayesian model to WSI task. They used the LDA-based model in which contexts of target word were viewed as documents as in the LDA model (Blei et al., 2003) and senses as topics. They trained a separate model for each target word and included a variety of features such as words, part-of-speech and dependency information. Yao and Van Durme (2011) extended LDA-based model into non-parametric HDP model but removed the feature engineering. Lau et al. (2012) showed improved supervised F-score by including position feature to the HDP model. Choe and Charniak (2013) proposed a reweighted naive Bayes model by incorporating the idea that words closer to the target word are more relevant in predicting the sense.

Our model differs from the context clustering models and graph-based models, as it is a Bayesian probabilistic model. Our work also differs from the LDA-based models. LDA topics were actually re-interpreted as senses of target word as Brody and Lapata (2009) applied the LDA to WSI tasks, so did Yao and Van Durme (2011) and Lau et al. (2012). They induced word senses by firstly tagging (sampling) senses (of target words) to context words and selecting the mostly tagged sense as sense of target words. Our model could be viewed as an extension of LDA, but fit the WSI task more naturally and much better. We distinguish senses of target words from concepts of context words and assume that they are separate. Therefore, our model has two hidden layers corresponding to the sense of the target word and the concepts of the context words respectively. Basically, one decide the sense of the target word based on the concept configuration of context words, instead of tagging senses of target word to context words. The separation of senses of target word and concepts of context words is actually not only required by linguistic intuition but also leads to improvement by our experiment. Our model is also different from the naive Bayes model since our model induces senses of the target word at concept level while naive Bayes model works at word level and does not involve conceptualization to context words at all.

7 Conclusion

In this paper, we first proposed a parametric Bayesian generative model to the task of Word Sense Induction. It is distinct from previous work in that it introduces a layer of latent concepts that generalize the context words and thus enable the model to measure the sense similarity at a more general level. We also show in this paper that Zipf's law of meaning can be used to guide the setting of sense numbers on an all-word basis, which is not only simple but also independent of the clustering methods being used. We further extend our parametric model to non-parametric model which not only simplifies the problem of model selection but also bring improved performance. The test results on the benchmark datasets show that our model outperforms the state-of-the-art systems.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant No. 61273318 and National Key Basic Research Program of China 2014CB340504.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 7–12. Association for Computational Linguistics.

Javier Artiles, Enrique Amigó, and Julio Gonzalo. 2009. The role of named entities in web people search. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 534–542. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–111. Association for Computational Linguistics.

George Casella and Edward I George. 1992. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Do Kook Choe and Eugene Charniak. 2013. Naive Bayes word sense induction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1437, Seattle, Washington, USA, October. Association for Computational Linguistics.

Phillip Edmonds. 2006. Disambiguation, lexical. *Encyclopedia of Language and Linguistics. Second Edition. Elsevier*.

Wesam Elshamy, Doina Caragea, and William H Hsu. 2010. Ksu kdd: Word sense induction by clustering in topic space. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 367–370. Association for Computational Linguistics.

Roman Kern, Markus Muhr, and Michael Granitzer. 2010. Kcdc: Word sense induction by using grammatical dependencies and sentence phrase structure. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 351–354. Association for Computational Linguistics.

Ioannis P Klapaftis and Suresh Manandhar. 2010. Word sense induction & disambiguation using hierarchical random graphs. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 745–755. Association for Computational Linguistics.

Ioannis Korkontzelos and Suresh Manandhar. 2010. Uoy: Graphs of unambiguous vertices for word sense induction and disambiguation. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 355–358. Association for Computational Linguistics.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601. Association for Computational Linguistics.

Jiwei Li and Claire Cardie. 2014. Timeline generation: Tracking individuals on twitter. In *Proceedings of the 23rd international conference on World wide web*, pages 643–652.

Jiwei Li and Sujian Li. 2013. Evolutionary hierarchical dirichlet process for timeline summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 556–560, Sofia, Bulgaria, August. Association for Computational Linguistics.

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

Zheng-Yu Niu, Dong-Hong Ji, and Chew-Lim Tan. 2007. I2r: Three systems for word sense discrimination, chinese word sense disambiguation, and english word sense disambiguation. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 177–182. Association for Computational Linguistics.

Ted Pedersen. 2010. Duluth-wsi: Senseclusters applied to the sense induction task of semeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 363–366. Association for Computational Linguistics.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, volume 7, pages 410–420.

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

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476).

Xuchen Yao and Benjamin Van Durme. 2011. Nonparametric bayesian word sense induction. In *Graph-based Methods for Natural Language Processing*, pages 10–14.

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