Knowledge-based Word Sense Disambiguation using Topic Models

Devendra Singh Chaplot, Ruslan Salakhutdinov
{chaplot,rsalakhu}@cs.cmu.edu
Machine Learning Department
School of Computer Science
Carnegie Mellon University

Abstract

Word Sense Disambiguation is an open problem in Natural Language Processing which is particularly challenging and useful in the unsupervised setting where all the words in any given text need to be disambiguated without using any labeled data. Typically WSD systems use the sentence or a small window of words around the target word as the context for disambiguation because their computational complexity scales exponentially with the size of the context. In this paper, we leverage the formalism of topic model to design a WSD system that scales linearly with the number of words in the context. As a result, our system is able to utilize the whole document as the context for a word to be disambiguated. The proposed method is a variant of Latent Dirichlet Allocation in which the topic proportions for a document are replaced by synset proportions. We further utilize the information in the WordNet by assigning a non-uniform prior to synset distribution over words and a logistic-normal prior for document distribution over synsets. We evaluate the proposed method on Senseval-2, Senseval-3, SemEval-2007, SemEval-2013 and SemEval-2015 English All-Word WSD datasets and show that it outperforms the state-of-the-art unsupervised knowledge-based WSD system by a significant margin.

Introduction

Word Sense Disambiguation (WSD) is the task of mapping an ambiguous word in a given context to its correct meaning. WSD is an important problem in natural language processing (NLP), both in its own right and as a stepping stone to more advanced tasks such as machine translation (Chan, Ng, and Chiang 2007), information extraction and retrieval (Zhong and Ng 2012), and question answering (Ramakrishnan et al. 2003). WSD, being AI-complete (Navigli 2009), is still an open problem after over two decades of research.

Following Navigli (2009), we can roughly distinguish between supervised and knowledge-based (unsupervised) approaches. Supervised methods require sense-annotated training data and are suitable for lexical sample WSD tasks where systems are required to disambiguate a restricted set of target words. However, the performance of supervised systems is limited in the all-word WSD tasks as labeled data for the full lexicon is sparse and difficult to obtain. As the all-word WSD task is more challenging and has more practical applications, there has been significant interest in developing unsupervised knowledge-based systems. These systems only require an external knowledge source (such as WordNet) but no labeled training data.

In this paper, we propose a novel knowledge-based WSD algorithm for the all-word WSD task, which utilizes the whole document as the context for a word, rather than just the current sentence used by most WSD systems. In order to model the whole document for WSD, we leverage the formalism of topic models, especially Latent Dirichlet Allocation (LDA). Our method is a variant of LDA in which the topic proportions for a document are replaced by synset proportions for a document. We use a non-uniform prior for the synset distribution over words to model the frequency of words within a synset. Furthermore, we also model the relationships between synsets by using a logistic-normal prior for drawing the synset proportions of the document. This makes our model similar to the correlated topic model , with the difference that our priors are not learned but fixed. In particular, the values of these priors are determined using the knowledge from WordNet. We evaluate our system on a set of five benchmark datasets, Senseval-2, Senseval-3, SemEval-2007, SemEval-2013 and SenEval-2015 and show that the proposed model outperforms state-of-the-art knowledge-based WSD system.

Related Work

Lesk (Lesk 1986) is a classical knowledge-based WSD algorithm which disambiguates a word by selecting a sense whose definition overlaps the most with the words in its context. Many subsequent knowledge-based systems are based on the Lesk algorithm. Banerjee and Pedersen (2003) extended Lesk by utilizing the definitions of words in the context and weighing the words by term frequency-inverse document frequency (tf-idf). Basile, Caputo, and Semeraro (2014) further enhanced Lesk by using word embeddings to calculate the similarity between sense definitions and words in the context.

The above methods only use the words in the context for disambiguating the target word. However, Chaplot, Bhattacharyya, and Paranjape (2015) show that sense of a word depends on not just the words in the context but also on their senses. Since the senses of the words in the context are also unknown, they need to be optimized jointly. In the

past decade, many graph-based unsupervised WSD methods have been developed which typically leverage the underlying structure of Lexical Knowledge Base such as WordNet and apply well-known graph-based techniques to efficiently select the best possible combination of senses in the context. Navigli and Lapata (2007) and Navigli and Lapata (2010) build a subgraph of the entire lexicon containing vertices useful for disambiguation and then use graph connectivity measures to determine the most appropriate senses. Mihalcea (2005) and Sinha and Mihalcea (2007) construct a sentence-wise graph, where, for each word every possible sense forms a vertex. Then graph-based iterative ranking and centrality algorithms are applied to find most probable sense. More recently, Agirre, López de Lacalle, and Soroa (2014) presented an unsupervised WSD approach based on personalized page rank over the graphs generated using WordNet. The graph is created by adding content words to the WordNet graph and connecting them to the synsets in which they appear in as strings. Then, the Personalized PageRank (PPR) algorithm is used to compute relative weights of the synsets according to their relative structural importance and consequently, for each content word, the synset with the highest PPR weight is chosen as the correct sense. Chaplot, Bhattacharyya, and Paranjape (2015) present a graph-based unsupervised WSD system which maximizes the total joint probability of all the senses in the context by modeling the WSD problem as a Markov Random Field constructed using the WordNet and a dependency parser and using a Maximum A Posteriori (MAP) Query for inference. Babelfy (Moro, Raganato, and Navigli 2014) is another graph-based approach which unifies WSD and Entity Linking (Rao, McNamee, and Dredze 2013). It performs WSD by performing random walks with restart over BabelNet (Navigli and Ponzetto 2012), which is a semantic network integrating WordNet with various knowledge resources.

The WSD systems which try to jointly optimize the sense of all words in the context have a common limitation that their computational complexity scales exponentially with the number of content words in the context due to pairwise comparisons between the content words. Consequently, practical implementations of these methods either use approximate or sub-optimal algorithms or reduce the size of context. Chaplot, Bhattacharyya, and Paranjape (2015) limit the context to a sentence and reduce the number of pairwise comparisons by using a dependency parser to extract important relations. Agirre, López de Lacalle, and Soroa (2014) limit the size of context to a window of 20 words around the target word. Moro, Raganato, and Navigli (2014) employ a densest subgraph heuristic for selecting high-coherence semantic interpretations of the input text. In contrast, the proposed approach scales linearly with the number of words in the context while optimizing sense of all the words in the context jointly. As a result, the whole document is utilized as the context for disambiguation.

Our work is also related to Boyd-Graber, Blei, and Zhu (2007) who were the first to apply LDA techniques to WSD. In their approach, words senses that share similar paths in the WordNet hierarchy are typically grouped in the same topic. However, they observe that WordNet is perhaps not

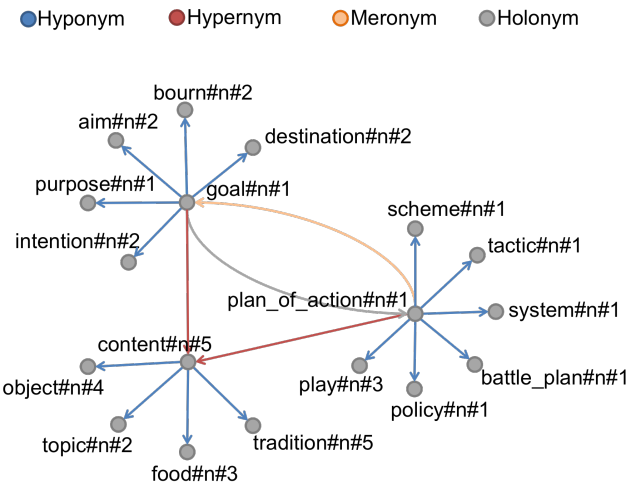

Figure 1: WordNet example showing several synsets and the relations between them.

the most optimal structure for WSD. Highly common, polysemous words such as man and time could potentially be associated with many different topics making decipherment of sense difficult. Even rare words that differ only subtly in their sense (e.g., quarterback – the position and quarterback – the player himself) could potentially only share the root node in WordNet and hence never have a chance of being on the same topic. Cai, Lee, and Teh (2007) also employ the idea of global context for the task of WSD using topic models, but rather than using topic models in an unsupervised fashion, they embed topic features in a supervised WSD model.

WordNet

Most WSD systems use a sense repository to obtain a set of possible senses for each word. WordNet is a comprehensive lexical database for the English language (Miller 1995), and is commonly used as the sense repository in WSD systems. It provides a set of possible senses for each content word (nouns, verbs, adjectives and adverbs) in the language and classifies this set of senses by the POS tags. For example, the word "cricket" can have 2 possible noun senses: 'cricket#n#1: leaping insect' and 'cricket#n#2: game played with a ball and bat', and a single possible verb sense, 'cricket#v#1: (play cricket)'. Furthermore, WordNet also groups words which share the same sense into an entity called synset (set of synonyms). Each synset also contains a gloss and example usage of the words present in it. For example, 'aim#n#2', 'object#n#2, 'objective#n#1', 'target#n#5' share a common synset having gloss "the goal intended to be attained".

WordNet also contains information about different types of semantic relationships between synsets. These relations include hypernymy, meronymy, hyponymy, holonymy, etc. Figure 1 shows a graph of a subset of the WordNet where nodes denote the synset and edges denote different semantic relationship between synsets. For instance, 'plan_of_action#n#1' is a meronym of 'goal#n#1', and 'pur-

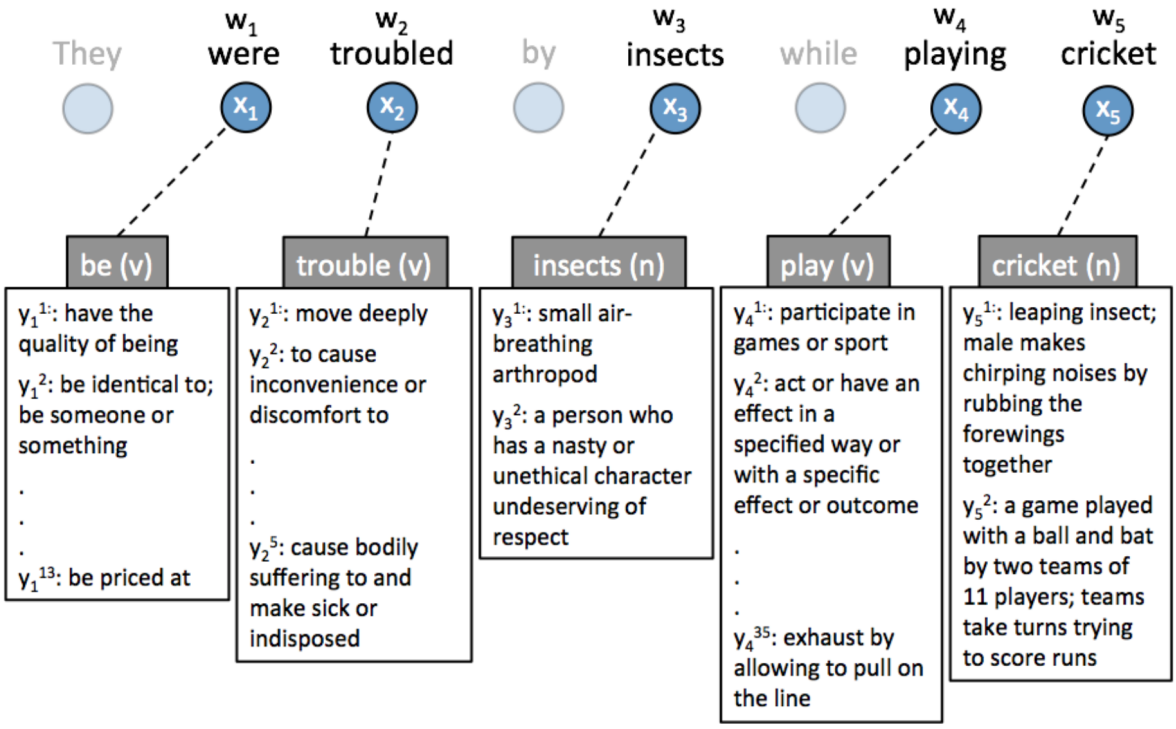

Figure 2: An example of the all-word WSD task. Content words and their possible senses are labeled w_i and y_i^j, respectively.

pose#n#1', 'aim#n#1' and 'destination#n#1' are hyponyms of 'goal#n#1', as shown in the figure. These semantic relationships in the WordNet can be used to compute the similarity between different synsets using various standard relatedness measures (Pedersen, Patwardhan, and Michelizzi 2004).

Methods

Problem Definition

First, we formally define the task of all-word Word Sense Disambiguation by illustrating an example. Consider a sentence, where we want to disambiguate all the content words (nouns, verbs, adjectives and adverbs):

They were troubled by insects while playing cricket.

The sense x_i of each content word (given its part-of-speech tag) w_i can take k_i possible values from the set $\boldsymbol{y}_i = \{y_i^1, y_i^2, \ldots, y_i^{k_i}\}$ (see Figure 2). In particular, the word $w_5 =$ "cricket" can either mean $y_5^1 =$ "a leaping insect" or $y_5^2 =$ "a game played with a ball and bat played by two teams of 11 players." In this example, the second sense is more appropriate. The problem of mapping each content word in any given text to its correct sense is called the all-word WSD task. The set of possible senses for each word is given by a sense repository like WordNet.

Semantics

In this subsection, we describe the semantic ideas underlying the proposed method and how they are incorporated in the proposed model:

- using the whole document as the context for WSD: modeled using Latent Dirichlet Allocation.
- some words in each synset are more frequent than others: modeled using non-uniform priors for the synset distribution over words.
- some synsets tend to co-occur more than others: modeled using logistic normal distribution for synset proportions in a document.

Wherever possible, we give examples to motivate the semantic ideas and illustrate their importance.

Document context The sense of a word depends on other words in its context. In the WordNet, the context of a word is defined to be the discourse that surrounds a language unit and helps to determine its interpretation. It is very difficult to determine the context of any given word. Most WSD systems use the sentence in which the word occurs as its context and each sentence is considered independent of others. However, we know that a document or an article is about a particular topic in some domain and all the sentences in a document are not independent of each other. Apart from the words in the sentence, the occurrence of certain words in a document might help in word sense disambiguation. For example, consider the following sentence,

He forgot the chips at the counter.

Here, the word 'chips' could refer to potato chips, micro chips or poker chips. It is not possible to disambiguate this word without looking at other words in the document. The presence of other words like 'casino', 'gambler', etc. in the document would indicate the sense of poker chips, while words like 'electronic' and 'silicon' indicate the sense

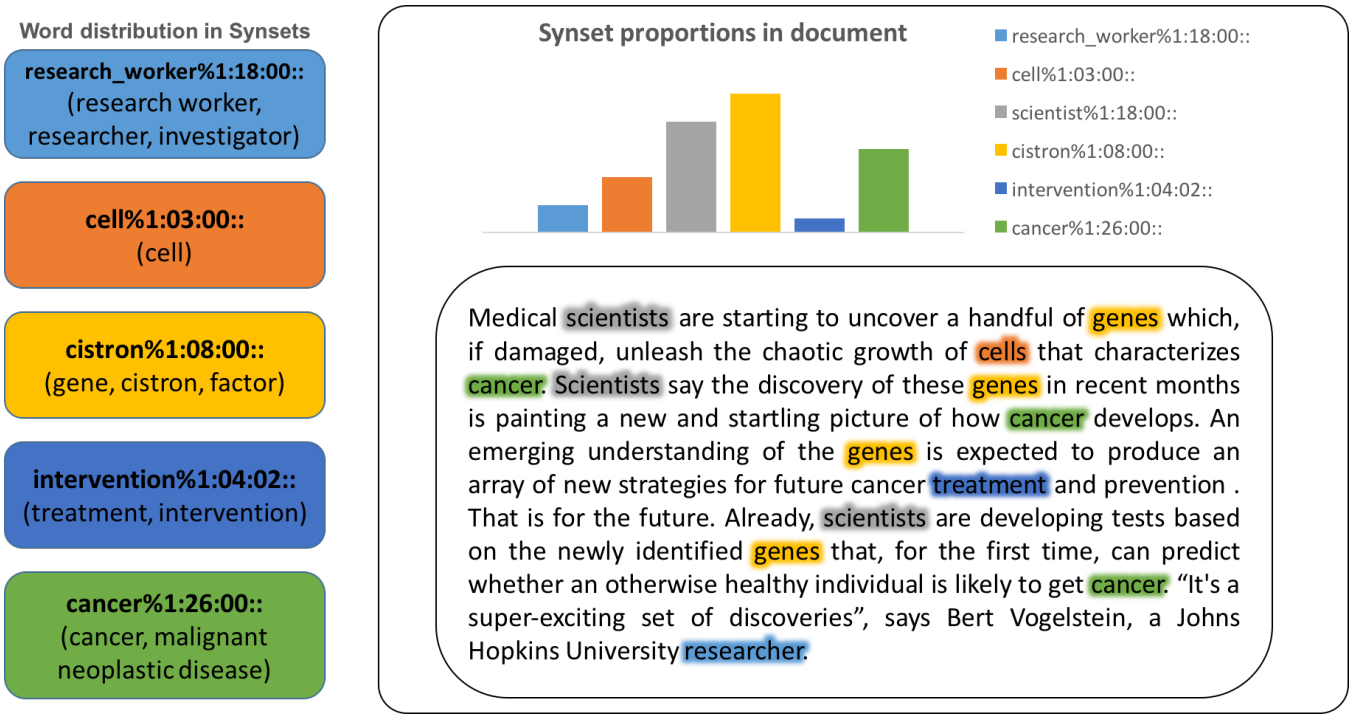

Figure 3: A toy example of word distribution in synsets and synset proportions in a document learned using the proposed model. Colors highlighting some of the words in the document denote the corresponding synsets they were sampled from.

of micro chip. Gale, Church, and Yarowsky (1992) also observed that words strongly tend to exhibit only one sense in a given discourse or document. Thus, we hypothesize that the meaning of the word depends on words outside the sentence in which it occurs – as a result, we use the whole document containing the word as its context.

Topic Models In order to use the whole document as the context for a word, we would like to model the concepts involved in the document. Topic models are suitable for this purpose, which aim to uncover the latent thematic structure in collections of documents (Blei 2012). The most basic example of a topic model is Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). It is based on the key assumption that documents exhibit multiple topics (which are nothing but distributions over some fixed vocabulary).

LDA has an implicit notion of word senses as words with several distinct meanings can appear in distinct topics (e.g., cricket the game in a "sports" topic and cricket the insect in a "zoology" topic). However, since the sense notion is only implicit (rather than a set of explicit senses for each word in WSD), it is not possible to directly apply LDA to the WSD task. Therefore, we modify the basic LDA by representing documents by synset probabilities rather than topic probabilities and consequently, the words are generated by synsets rather than topics. We further modify this graphical model to incorporate the information in the WordNet as described in the following subsections.

Synset distribution over words Due to sparsity problems in large vocabulary size, the LDA model was extended to a "smoothed" LDA model by placing an exchangeable Dirichlet prior on topic distribution over words. In an exchangeable Dirichlet distribution, each component of the parameter vector is equal to the same scalar. However, such a uniform prior is not ideal for synset distribution over words since each synset contains only a fixed set of words. For example, the synset defined as "the place designated as the end (as of a race or journey)" contains only 'goal','destination' and 'finish'. Furthermore, some words in each synset are more frequent than others. For example, in the synset defined as "a person who participates in or is skilled at some game", the word 'player' is more frequent than word 'participant', while in the synset defined as "a theatrical performer", word 'actor' is more frequent than word 'player'. Thus, we decide to have non-uniform priors for synset distribution over words.

Document distribution over synsets The LDA model uses a Dirichlet distribution for the topic proportions of a document. Under a Dirichlet, the components of the topic proportions vector are approximately independent; this leads to the strong and unrealistic modeling assumption that the presence of one topic is not correlated with the presence of other topics. Similarly, in our case, the presence of one synset is correlated with the presence of others. For example, the synset representing the 'sloping land' sense of the word 'bank' is more likely to cooccur with the synset of 'river' (a large natural stream of water) than the synset represent-

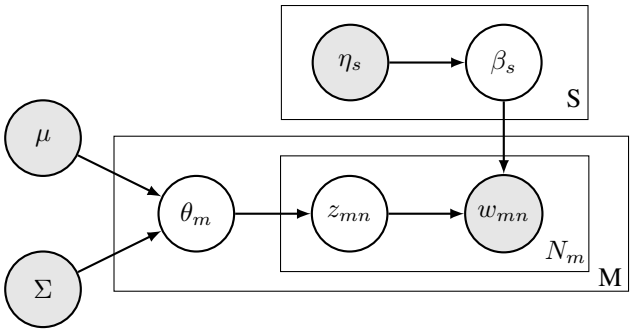

Figure 4: Graphical model for the proposed method.

ing 'financial institution' sense of the word 'bank'. Hence, we model the correlations between synsets using a logistic normal distribution for synset proportions in a document.

Proposed Model

Following the ideas described in the previous subsection, we propose a probabilistic graphical model, which assumes that a corpus is generated according to the following process:

1. For each synset, $s \in \{1, \ldots, S\}$
   (a) Draw word proportions $\beta_s \sim \text{Dir}(\eta_s)$
2. For each document, $m \in \{1, \ldots, M\}$
   (a) Draw $\alpha_m \sim \mathcal{N}(\mu, \Sigma)$
   (b) Transform α_m to synset proportions $\theta_m = f(\alpha_m)$
   (c) For each word in the document, $n \epsilon \{1, \ldots, N_m\}$
      i. Draw synset assignment $z_{mn} \sim \text{Mult}(\theta_m)$
      ii. Draw word from assigned synset $w_{mn} \sim \text{Mult}(\beta_{z_{mn}})$

where $f(\boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{\alpha})}{\sum_i \exp(\alpha_i)}$ is the softmax function.

Note that the prior for drawing word proportions for each sense is not symmetric: η_s is a vector of length equal to word vocabulary size, having non-zero equal entries only for the words contained in synset s in WordNet. The graphical model corresponding to the generative process is shown in Figure 4. Figure 3 illustrates a toy example of a possible word distribution in synsets and synset proportions in a document learned using the proposed model. Colors highlighting some of the words in the document denote the corresponding synsets they were sampled from.

Priors

We utilize the information in the WordNet for deciding the priors for drawing the word proportions for each synset and the synset proportions for each document. The prior for distribution of synset s over words is chosen as the frequency of the words in the synset s, i.e.,

$$\eta_{sv} = \text{Frequency of word } v \text{ in synset } s.$$

The logistic normal distribution for drawing synset proportions has two priors, μ and Σ. The parameter μ_s gives the probability of choosing a synset s. The frequency of the synset s would be the natural choice for μ_s but since our method is unsupervised, we use a uniform μ for all synsets instead. The Σ parameter is used to model the relationship between synsets. Since, the inverse of covariance matrix will be used in inference, we directly choose (i, j)th element of inverse of covariance matrix as follows:

Σ^{-1}_{ij} = Negative of similarity between synset i and synset j

The similarity between any two synsets in the WordNet can be calculated using a variety of relatedness measures given in WordNet::Similarity library (Pedersen, Patwardhan, and Michelizzi 2004). In this paper, we use the Lesk similarity measure as it is used in prior WSD systems. Lesk algorithm calculates the similarity between two synsets using the overlap between their definitions.

Inference

We use a Gibbs Sampler for sampling latent synsets z_{mn} given the values of rest of the variables. Given a corpus of M documents, the posterior over latent variables, i.e. the synset assignments $\boldsymbol{z}$, logistic normal parameter $\boldsymbol{\alpha}$, is as follows:

$$\begin{aligned} & p(\boldsymbol{z}, \boldsymbol{\alpha} | \boldsymbol{w}, \boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ & \propto \quad p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta})\ p(\boldsymbol{\beta}|\boldsymbol{\eta})\ p(\boldsymbol{z}|\boldsymbol{\alpha})\ p(\boldsymbol{\alpha}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \end{aligned}$$

The word distribution $p(w_{mn}|z_{mn}, \beta)$ is multinomial in $\beta_{z_{mn}}$ and the conjugate distribution $p(\beta_s|\eta_s)$ is Dirichlet in η_s. Thus, $p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta})\ p(\boldsymbol{\beta}|\boldsymbol{\eta})$ can be collapsed to $p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\eta})$ by integrating out β_s for all senses s to obtain:

$$p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\eta}) = \prod_{s=1}^{S} \frac{\prod_v \Gamma(n^{SV}_{sv} + \eta_{sv})}{\Gamma(n^S_s + ||\eta_s||_1)} \frac{\Gamma(||\eta_s||_1)}{\prod_s \Gamma(\eta_{sv})}$$

The sense distribution $p(z_{mn}|\boldsymbol{\alpha_m})$ is a multinomial distribution with parameters $f(\boldsymbol{\alpha_m}) = \boldsymbol{\theta_m}$:

$$p(\boldsymbol{z}|\boldsymbol{\alpha}) = \prod_{m=1}^{M} \left(\prod_{n=1}^{N_m} \frac{exp(\alpha_m^{z_{mn}})}{\sum_{s=1}^{S} exp(\alpha_m^s)} \right)$$

$\boldsymbol{\alpha_m}$ follows a normal distribution which is not conjugate of the multinomial distribution:

$$p(\boldsymbol{\alpha_m}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \mathcal{N}(\boldsymbol{\alpha_m}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Thus, $p(\boldsymbol{z}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can't be collapsed. In typical logistic-normal topic models, a block-wise Gibbs sampling algorithm is used for alternatively sampling topic assignments and logistic-normal parameters. However, since in

	System	Senseval-2	Senseval-3	SemEval-07	SemEval-13	SemEval-15	All
Knowledge based	Banerjee03	50.6	44.5	32.0	53.6	51.0	48.7
	Basile14	63.0	63.7	**56.7**	66.2	64.6	63.7
	Agirre14	60.6	54.1	42.0	59.0	61.2	57.5
	Moro14	67.0	63.5	51.6	**66.4**	**70.3**	65.5
	WSD-TM	**69.0**	**66.9**	55.6	65.3	69.6	**66.9**
Supervised	MFS	66.5	60.4	52.3	62.6	64.2	62.9
	Zhong10	70.8	68.9	58.5	66.3	69.7	68.3
	Melamud16	72.3	68.2	61.5	67.2	71.7	69.4

Table 1: Comparison of F1 scores with various WSD systems on English all-words datasets of Senseval-2, Senseval-3, SemEval-2007, SemEval-2013, SemEval-2015. WSD-TM corresponds to the proposed method. The best results in each column among knowledge-based systems are marked in **bold**.

our case the logistic-normal priors are fixed, we can sample synset assignments directly using the following equation:

$$\begin{aligned} & p(z_{mn} = k|rest) \\ = \quad & \frac{p(z, w|\alpha, \eta)}{p(z_{-mn}, w|\alpha, \eta)} \\ \propto \quad & p(z, w|\alpha, \eta) \\ \propto \quad & \frac{(\eta_{sv} + n^{SV}_{sv_{-mn}})}{n^{S}_{s_{-mn}} + ||\eta_s||_1} \exp(\alpha^k_m) \end{aligned}$$

Here, n^{SV}_{sv}, n^{SM}_{sm} and n^{S}_{s} correspond to standard counts:

$$\begin{aligned} n^{SV}_{sv} &= \sum_{m,n} \{z_{mn} = s, w_{mn} = v\} \\ n^{SM}_{sm} &= \sum_{n} \{z_{mn} = s\} \\ n^{S}_{s} &= \sum_{m} n^{SM}_{sm} \end{aligned}$$

The subscript $\cdot_{-mn}$ denotes the corresponding count without considering the word n in document m. The value of z_{mn} at the end of Gibbs sampling is the labelled sense of word n in document m. The computational complexity of Gibbs sampling for this graphical model is linear with respect to number of words in the document (Qiu et al. 2014).

Experiments & Results

For evaluating our system, we use the English all-word WSD task benchmarks of the SensEval-2 (Palmer et al. 2001), SensEval-3 (Snyder and Palmer 2004), SemEval-2007 (Pradhan et al. 2007), SemEval-2013 (Navigli, Jurgens, and Vannella 2013) and SemEval-2015 (Moro and Navigli 2015). (Raganato, Camacho-Collados, and Navigli 2017) standardized all the above datasets into a unified format with gold standard keys in WordNet 3.0. We use the standardized version of all the datasets and use the same experimental setting as (Raganato, Camacho-Collados, and Navigli 2017) for fair comparison with prior methods.

In Table 1 we compare our overall F1 scores with different unsupervised systems described in Section 2 which include Banerjee03 (Banerjee and Pedersen 2003), Basile14 (Basile, Caputo, and Semeraro 2014), Agirre14 (Agirre, López de Lacalle, and Soroa 2014) and Moro14 (Moro, Raganato, and Navigli 2014). In addition to knowledge-based systems which do not require any labeled training corpora, we also report F1 scores of the state-of-the-art supervised systems trained on SemCor (Miller et al. 1994) and OMSTI (Taghipour and Ng 2015) for comparison. Zhong10 (Zhong and Ng 2010) use a Support Vector Machine over a set of features which include surrounding words in the context, their PoS tags, and local collocations. Melmaud16 (Melamud, Goldberger, and Dagan 2016) learn context embeddings of a word and classify a test word instance with the sense of the training set word whose context embedding is the most similar to the context embedding of the test instance. We also provide the F1 scores of MFS baseline, i.e. labeling each word with its most frequent sense (MFS) in labeled datasets, SemCor (Miller et al. 1994) and OMSTI (Taghipour and Ng 2015).

The proposed method, denoted by WSD-TM in the tables referring to WSD using topic models, outperforms the state-of-the-art WSD system by a significant margin (p-value $<$0.01) by achieving an overall F1-score of 66.9 as compared to Moro14's score of 65.5. We also observe that the performance of the proposed model is not much worse than the best supervised system, Melamud16 (69.4). In Table 2 we report the F1 scores on different parts of speech. The proposed system outperforms all previous knowledge-based systems over all parts of speech. This indicates that using document context helps in disambiguating words of all PoS tags.

Discussions

In this section, we illustrate the benefit of using the whole document as the context for disambiguation by illustrating an example. Consider an excerpt from the SensEval 2 dataset shown in Figure 3. Highlighted words clearly indicate that the domain of the document is Biology. While most of these words are monosemous, let's consider disambiguating the word 'cell', which is highly polysemous, having 7 possible senses as shown in Figure 5. As shown in Table 3, the correct sense of cell ('cell#2') has the highest similarity with senses of three monosemous words 'scientist', 'researcher' and 'protein'. The word 'cell' occurs 21 times in

	System	Nouns	Verbs	Adjectives	Adverbs	All
Knowledge based	Banerjee03	54.1	27.9	54.6	60.3	48.7
	Basile14	69.8	51.2	51.7	80.6	63.7
	Agirre14	62.1	38.3	66.8	66.2	57.5
	Moro14	68.6	49.9	73.2	79.8	65.5
	WSD-TM	**69.7**	**51.2**	**76.0**	**80.9**	**66.9**
Supervised	MFS	65.8	45.9	72.7	80.5	62.9
	Zhong10	71.0	53.3	77.1	82.7	68.3
	Melamud16	71.7	55.8	77.2	82.7	69.4

Table 2: Comparison of F1 scores on different POS tags over all datasets. WSD-TM corresponds to the proposed method. The best results in each column among knowledge-based systems are marked in **bold**.

Sense of 'cell'	Similarity with		
	scientist#1	researcher#1	protein#1
cell#1	0.100	0.091	0.077
cell#2	**0.200**	**0.167**	**0.100**
cell#3	0.100	0.091	0.077
cell#4	0.100	0.062	0.071
cell#5	0.100	0.077	0.067
cell#6	0.100	0.091	0.077
cell#7	0.100	0.091	0.077

Table 3: The similarity of different senses of the word 'cell' with senses of three monosemous words 'scientist', 'researcher' and 'protein'. The correct sense of cell, 'cell#2', has the highest similarity with all the three synsets.

- S: (n) **cell#1** (any small compartment) *"the cells of a honeycomb"*
- S: (n) **cell#2** ((biology) the basic structural and functional unit of all organisms; they may exist as independent units of life (as in monads) or may form colonies or tissues as in higher plants and animals)
- S: (n) **cell#3**, electric cell#1 (a device that delivers an electric current as the result of a chemical reaction)
- S: (n) **cell#4**, cadre#1 (a small unit serving as part of or as the nucleus of a larger political movement)
- S: (n) cellular telephone#1, cellular phone#1, cellphone#1, **cell#5**, mobile phone#1 (a hand-held mobile radiotelephone for use in an area divided into small sections, each with its own short-range transmitter/receiver)
- S: (n) **cell#6**, cubicle#1 (small room in which a monk or nun lives)
- S: (n) **cell#7**, jail cell#1, prison cell#1 (a room where a prisoner is kept)

Figure 5: The different senses of the word 'cell'

the document, and several times, the other words in the sentence are not adequate to disambiguate it. Since our method uses the whole document as the context, words such as 'scientists', 'researchers' and 'protein' help in disambiguating 'cell', which is not possible otherwise.

The proposed model also overcomes several limitations of topic models based on Latent Dirichlet Allocation and its variants. Firstly, LDA requires the specification of the number of topics as a hyper-parameter which is difficult to tune. The proposed model doesn't require total number of synsets to be specified as the total number of synsets are equal to the number of synsets in the sense repository which is fixed. Secondly, topics learned using LDA are often not meaningful as the words inside some topics are unrelated. However, synsets are always meaningful as they contain only synonymous words.

Conclusion

In this paper, we propose a novel knowledge-based WSD system based on a logistic normal topic model which incorporates semantic information about synsets as its priors. The proposed model scales linearly with the number of words in the context, which allows our system to use the whole document as the context for disambiguation and outperform state-of-the-art knowledge-based WSD system on a set of benchmark datasets.

One possible avenue for future research is to use this model for supervised WSD. This could be done by using sense tags from the SemCor corpus as training data in a supervised topic model similar to the one presented by (Mcauliffe and Blei 2008). Another possibility would be to add another level to the hierarchy of the document generating process. This would allow us to bring back the notion of topics and then to define topic-specific sense distributions. The same model can also be extended to other problems such named-entity disambiguation.

References

[Agirre, López de Lacalle, and Soroa 2014] Agirre, E.; López de Lacalle, O.; and Soroa, A. 2014. Random walks for knowledge-based word sense disambiguation. *Comput. Linguist.* 40(1):57–84.

[Banerjee and Pedersen 2003] Banerjee, S., and Pedersen, T. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Ijcai*, volume 3, 805–810.

[Basile, Caputo, and Semeraro 2014] Basile, P.; Caputo, A.; and Semeraro, G. 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *COLING*, 1591–1600.

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

[Blei 2012] Blei, D. M. 2012. Probabilistic topic models. *Communications of the ACM* 55(4):77–84.

[Boyd-Graber, Blei, and Zhu 2007] Boyd-Graber, J. L.; Blei, D. M.; and Zhu, X. 2007. A topic model for word sense disambiguation. In *EMNLP-CoNLL*, 1024–1033.

[Cai, Lee, and Teh 2007] Cai, J.; Lee, W. S.; and Teh, Y. W. 2007. Improving word sense disambiguation using topic features. In *EMNLP-CoNLL*, 1015–1023.

[Chan, Ng, and Chiang 2007] Chan, Y. S.; Ng, H. T.; and Chiang, D. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 33–40. Prague, Czech Republic: Association for Computational Linguistics.

[Chaplot, Bhattacharyya, and Paranjape 2015] Chaplot, D. S.; Bhattacharyya, P.; and Paranjape, A. 2015. Unsupervised word sense disambiguation using markov random field and dependency parser. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[Gale, Church, and Yarowsky 1992] Gale, W. A.; Church, K. W.; and Yarowsky, D. 1992. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, 233–237. Association for Computational Linguistics.

[Lesk 1986] Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, 24–26. ACM.

[Mcauliffe and Blei 2008] Mcauliffe, J. D., and Blei, D. M. 2008. Supervised topic models. In *Advances in neural information processing systems*, 121–128.

[Melamud, Goldberger, and Dagan 2016] Melamud, O.; Goldberger, J.; and Dagan, I. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *CoNLL*, 51–61.

[Mihalcea 2005] Mihalcea, R. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 411–418. Association for Computational Linguistics.

[Miller et al. 1994] Miller, G. A.; Chodorow, M.; Landes, S.; Leacock, C.; and Thomas, R. G. 1994. Using a semantic concordance for sense identification. In *Proceedings of the workshop on Human Language Technology*, 240–243. Association for Computational Linguistics.

[Miller 1995] Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

[Moro and Navigli 2015] Moro, A., and Navigli, R. 2015. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *SemEval@ NAACL-HLT*, 288–297.

[Moro, Raganato, and Navigli 2014] Moro, A.; Raganato, A.; and Navigli, R. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics* 2:231–244.

[Navigli and Lapata 2007] Navigli, R., and Lapata, M. 2007. Graph connectivity measures for unsupervised word sense disambiguation. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, 1683–1688. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[Navigli and Lapata 2010] Navigli, R., and Lapata, M. 2010. An study of graph connectivity for unsupervised word sense disambiguation. *IEEE transactions on pattern analysis and machine intelligence* 32(4):678–692.

[Navigli and Ponzetto 2012] Navigli, R., and Ponzetto, S. P. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193:217–250.

[Navigli, Jurgens, and Vannella 2013] Navigli, R.; Jurgens, D.; and Vannella, D. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *SemEval@ NAACL-HLT*, 222–231.

[Navigli 2009] Navigli, R. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)* 41(2):10.

[Palmer et al. 2001] Palmer, M.; Fellbaum, C.; Cotton, S.; Delfs, L.; and Dang, H. T. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, 21–24. Toulouse, France: Association for Computational Linguistics.

[Pedersen, Patwardhan, and Michelizzi 2004] Pedersen, T.; Patwardhan, S.; and Michelizzi, J. 2004. Wordnet::similarity: Measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, 38–41. Stroudsburg, PA, USA: Association for Computational Linguistics.

[Pradhan et al. 2007] Pradhan, S. S.; Loper, E.; Dligach, D.; and Palmer, M. 2007. Semeval-2007 task 17: English lexical sample, srl and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, 87–92. Association for Computational Linguistics.

[Qiu et al. 2014] Qiu, Z.; Wu, B.; Wang, B.; Shi, C.; and Yu, L. 2014. Collapsed gibbs sampling for latent dirichlet allocation on spark. In *Proceedings of the 3rd International Conference on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications-Volume 36*, 17–28. JMLR. org.

[Raganato, Camacho-Collados, and Navigli 2017] Raganato, A.; Camacho-Collados, J.; and Navigli, R. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proc. of EACL*, 99–110.

[Ramakrishnan et al. 2003] Ramakrishnan, G.; Jadhav, A.; Joshi, A.; Chakrabarti, S.; and Bhattacharyya, P. 2003. Question answering via bayesian inference on lexical relations. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering - Volume 12*, 1–10. Stroudsburg, PA, USA: Association for Computational Linguistics.

[Rao, McNamee, and Dredze 2013] Rao, D.; McNamee, P.; and Dredze, M. 2013. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*. Springer. 93–115.

[Sinha and Mihalcea 2007] Sinha, R., and Mihalcea, R. 2007. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In *Proceedings of the International Conference on Semantic Computing*, 363–369. Washington, DC, USA: IEEE Computer Society.

[Snyder and Palmer 2004] Snyder, B., and Palmer, M. 2004. The english all-words task. In Mihalcea, R., and Edmonds, P., eds., *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 41–43. Barcelona, Spain: Association for Computational Linguistics.

[Taghipour and Ng 2015] Taghipour, K., and Ng, H. T. 2015. One million sense-tagged instances for word sense disambiguation and induction. *CoNLL 2015* 338.

[Zhong and Ng 2010] Zhong, Z., and Ng, H. T. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, 78–83. Association for Computational Linguistics.

[Zhong and Ng 2012] Zhong, Z., and Ng, H. T. 2012. Word sense disambiguation improves information retrieval. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, 273–282. Stroudsburg, PA, USA: Association for Computational Linguistics.