Multi Sense Embeddings from Topic Models

Shobhit Jain
Amazon Web Services
jainshob@amazon.com

Sravan Babu Bodapati
Amazon Web Services
sravanb@amazon.com

Ramesh Nallapati
Amazon Web Services
rnallapa@amazon.com

Abstract

Distributed word embeddings have yielded state-of-the-art performance in many NLP tasks, mainly due to their success in capturing useful semantic information. These representations assign only a single vector to each word whereas a large number of words are polysemous (i.e., have multiple meanings). In this work, we approach this critical problem in lexical semantics, namely that of representing various senses of polysemous words in vector spaces. We propose a topic modeling based skip-gram approach for learning multi-prototype word embeddings. We also introduce a method to prune the embeddings determined by the probabilistic representation of the word in each topic. We use our embeddings to show that they can capture the context and word similarity strongly and outperform various state-of-the-art implementations.

1 Introduction

Representing words as dense, low dimensional embeddings (Mikolov et al., 2013a,b; Pennington et al., 2014) allow the representations to capture useful syntactic & semantic information making them useful in downstream Natural Language Processing tasks. However, these embedding models ignore the lexical ambiguity among different meanings of the same word. They assign only a single vector representative of all the different meanings of a word. In this work, we attempt to address this problem by capturing the multiple senses of a word using the global semantics of the document in which the word appears. Li and Jurafsky (2015) indicated that such sense specific vectors improve the performance of applications related to semantic understanding, such as Named Entity Recognition, Part-Of-Speech tagging.

In this work, we first train a topic model on our corpus to extract the topic distribution for each document. We treat these extracted topics as a heuristic to model word senses. We hypothesize that these word senses correlate quite well with the human notion of word senses, and validate it through our rigorous experiments as we demonstrate in our results section. We then use this topic distribution to train sense-specific word embeddings for each sense. We train these embeddings by weighing the learning procedure in proportion to the corresponding topic representation for each document. However, a word need not strongly correlate with each of these extracted senses. To address it, we propose a variant of this model which restricts the learning to only those embeddings where the word has a strong correlation with the topic extracted, i.e., high $p(word|topic)$.

The major contributions of our work are (i) training multi-sense word embeddings based on structured skip gram using topic models as a precursor (ii) non-parametric approach which prunes the embeddings to capture variability in the number of word senses.

2 Prior Work

Recently, learning multi-sense word embedding models has been an active area of research and has gained a lot of interest. TF-IDF (Reisinger and Mooney, 2010), SaSA (Wu and Giles, 2015), MSSG (Neelakantan et al., 2015), Huang et al. (2012) used cluster-based techniques to cluster the context of a word and comprehend word senses from the cluster centroids. Tian et al. (2014) proposed to use EM-based probabilistic clustering to assign word senses. Li and Jurafsky (2015) used Chinese Restaurant Process to model the word senses. All these techniques are just local context based and thus ignore the essential correlations amongst words and phrases in a broader document-level context. In contrast, our method enriches the embeddings with the document level information,

capturing word interactions in a broader document-level context.

AutoExtend (Rothe and Schütze, 2015), Sensembed (Iacobacci et al., 2015), Nasari (Camacho-Collados et al., 2016), Deconf (Pilehvar and Collier, 2016), Chen et al. (2014); Jauhar et al. (2015); Pelevina et al. (2017) have used multi-step approach to learn sense & word embeddings but require an external lexical database like WordNet to achieve it. SW2V(Mancini et al., 2016) train the embeddings in a single joint training phase. Nonetheless, all these methods assign same weight to every sense of a word, ignoring the extent to which each sense is associated with it's context.

MSWE (Nguyen et al., 2017) trained sense and word embeddings separately, with sense specific word embeddings computed as a weighted sum of the two, where the weights are calculated using topic modeling. Similarly, Liu et al. (2015a,b); Cheng et al. (2015); Zhang and Zhong (2016) used skip-gram based approach to obtain separate word & topic embeddings. Lau et al. (2013) also used topic models to distinguish between different senses of a word. All these techniques express the sense-specific word representation as a function of word & sense embeddings which essentially belongs to two different domains. Our work trains more robust compositional word embeddings formulated as a weighted sum of sense specific word embeddings, thus, taking into consideration all the different word senses while operating in the same vector space.

More recent techniques like ELMo (Peters et al., 2018), BERT (Devlin et al., 2018) compute the contextual representations of a word based on the sentence in which the word appears, whereas, our method yields precomputed embeddings for each sense of a word within the same vector space.

3 Multi Sense Embeddings Model

3.1 Topic Modeling

Mixed membership models like topic models allow us to discover topics that occur in a collection of documents. A *topic* is defined as a distribution over words and consists of cluster of words that occur frequently. This formulation benefits us in inferring the probability distribution over different contexts(topics) the word can occur in. Latent Dirichlet Allocation(LDA) (Blei et al., 2003) is a topic modeling technique that assigns multiple topics in different proportions to each document along with the probability distribution over words for each of the topics. Topic models based on Gibbs Sampling (Geman and Geman, 1987) achieve this by computing the posterior for a word based on the topic proportion at document level coupled with how often the word appears together with other words in the topic. We use Gibbs Sampling based approach to compute the topic distribution for each document. We use the LDA implementation from MALLET *topic modeling* toolkit (McCallum, 2002) for our experiments.

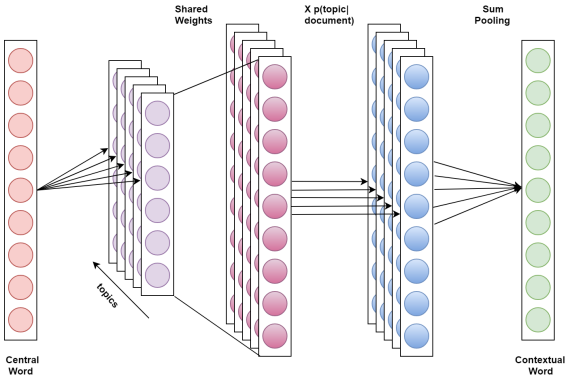

Figure 1: We feed our model a central word as input and predict the context word from it. First we learn separate word embeddings corresponding to each topic, E_{w_t,z_i}. Each of these embeddings are then multiplied with global word embeddings, $v_g(w)$, weighed in proportion to the topic distribution of the document from which the words have been chosen, and summed up to predict the neighboring context word.

3.2 Embeddings from Topic Models (ETMo)

In this section we present our baseline approach for training sense-specific word embeddings. We formulate our approach as follows. Let $E_w \in R^{k \times n}$ represent the embedding matrix for word w, where k is the number of topics(treated as number of word senses) and n is the dimensionality of embeddings. We represent the embedding of word w corresponding to topic z_i as E_{w,z_i}. Let $v_g(w)$ be the *output* vector representation for word w, which is shared across senses, and enforces the embeddings of different senses to be within the same vector space.

We introduce a latent variable z, representing the topic dimension, to model separate embedding for each topic. Inline with the skip-gram(Mikolov et al., 2013a) approach, we maximize the probability of predicting the context word w_{t+j}, given a

central word w_t for a document d as:

$$p(w_{t+j}|w_t, d) = \sum_{i=1}^{k} p(w_{t+j}|w_t, z_i, d) * p(z_i|d) \quad (1)$$

$p(z_i|d)$ represents the topic distribution of the document d, obtained from the trained topic model. In the above equation, we reasonably make the assumption, $p(z_i|w_t, d) = p(z_i|d)$, owing to the fact that the topic distribution is computed at the document level. Using Negative Sampling (Mikolov et al., 2013b), we reduce the first term in the above equation as:

$$p(w_{t+j}|w_t, z_i) = \sigma(E_{w_t,z_i} * v_g(w_{t+j})) + \sum_{w \in S} \sigma(-E_{w_t,z_i} * v_g(w)) \quad (2)$$

Formally, given a large corpus of documents, with size D, having a words sequence $w_1, w_2, ..., w_{N_d-1}, w_{N_d}$, where N_d is the number of words in document d, skip-window size c, number of topics k, the objective is to maximize the following log likelihood:

$$L = \sum_{d=1}^{D} \sum_{t=1}^{N_d} \sum_{j=-c}^{c} log\ p(w_{t+j}|w_t, d) = \sum_{d=1}^{D} \sum_{t=1}^{N_d} \sum_{j=-c}^{c} log \sum_{i=1}^{k} p(w_{t+j}|w_t, z_i, d) * p(z_i|d) \quad (3)$$

As shown in Figure 1, we use a neural network architecture to compute the log likelihood. We feed the central word, in its BoW representation, as input to the model and compute the probability of the context word. Refer to the figure for detailed explanation.

During inference, we first compute the topic distribution for the given document, $p(z_i|d)$, using our pre-trained topic model. Finally, for a document d and for each word w, we infer the word embedding as:

$$v_{w,d} = \sum_{i=1}^{k} p(z_i|d) * E_{w,z_i} \quad (4)$$

Model	$avgSim$	$globalSim$
GloVe (Pennington et al., 2014)	-	63.2
Huang et al. (2012)	64.2	71.3
csmRNN (Luong et al., 2013)	-	64.58
GC-SINGLE (Jauhar et al., 2015)	62.3	-
NP-MSSG (Neelakantan et al., 2015)	69.1	68.6
MSWE-I (Nguyen et al., 2017)	-	**72.40**
Gensense (Lee et al., 2018)	54.0	-
ETMo (Ours)	68.5	68.2
ETMo + NP (Ours)	**69.3**	69.1

Table 1: Spearman's correlation $\rho \times 100$ on WS-353

3.3 ETMo + Non-parametric

In this section, we substantiate the flaws in our baseline approach and present our non-parametric method to learn the embeddings.

Our previous approach assigns an embedding to every word corresponding to each topic. As one can see, this method would undesirably accumulate a fair amount of noisy updates to those word embeddings that have minimal representation in a topic. Hence, we extend our model by exploiting the information from topic models to learn only those embeddings where the word has a strong correlation with the topic.

In particular, we train only those embedding E_{w_t,z_i} such that $p(w_t|z_i) > p_{thres}$, where p_{thres} is chosen empirically, which we will explain later. For the words where none of the senses satisfy the above condition (might be the case for some monosemous words), we chose the embedding $E_{w_t,x}$ to be trained, such that $x = \text{argmax}_{z_i}\ p(w_t|z_i)$.

4 Experimental Setup

We use the English Wikipedia corpus dump (Shaoul and Westbury, 2010) for training both, topic models and embedding models. Though many previous research works have used a larger training corpus, but for a fair comparison, we only compare our results with those works which have used the same corpus. We could also improve obtained results by using a larger training corpus, but this is not central point of our paper. The main aim of our work is to compute sense specific embed-

Model	$avgSim$	$avgSimC$
TF-IDF	60.4	-
Huang et al. (2012)	62.8	65.7
Tian et al. (2014)	-	65.4
Chen et al. (2014)	66.2	68.9
Cheng et al. (2015)	-	65.9
GC-MULTI (Jauhar et al., 2015)	-	65.9
SENSEMBED (Iacobacci et al., 2015)	62.4	-
SaSA (Wu and Giles, 2015)	-	66.4
TWE-I (Liu et al., 2015b)	-	68.1
NP-MSSG (Neelakantan et al., 2015)	67.2	**69.3**
SG+Greedy (Li and Jurafsky, 2015)	-	69.1
MSWE (Nguyen et al., 2017)	66.7	66.6
ETMo (Ours)	65.4	65.8
ETMo + NP (Ours)	**67.5**	69.1

Table 2: Spearman's correlation $\rho \times 100$ on SCWS

dings for a word using topic models and demonstrate the strength of our model empirically.

The raw dataset consists of nearly 3.2 million documents and 1 billion tokens. Training topic models on such a large and diverse corpus helps in obtaining clearly demarcated senses for each topic.

To tune the hyper parameters of our neural network model, we sample 20% of our corpus as validation data and chose those parameters that give the lowest validation loss. Later, we use these parameter values for training on the entire corpus. For all our experiments, we use a skip-window of size 2, 8 negative samples, embeddings of dimensionality 200, and fix the number of topics to 10. A detailed analysis on how we chose the number of topics, using perplexity score, can be found later in the analysis section. We initialize the embeddings using pre-trained GloVe embeddings to ensure all our target embeddings are in the same vector space. We choose the value for p_{thres} as 1e-4 and give an analysis on how we chose the parameter value in the results section.

5 Results

We evaluate our model on two tasks, namely, word similarity and word analogy. For word similarity evaluation, we evaluate our embeddings on standard word similarity benchmark datasets including WS-353 (Finkelstein et al., 2001) & SCWS-2003 (Huang et al., 2012). WS-353 includes 353 pairs

Model	Accuracy(%)
Word2Vec	67
Huang et al. (2012)	12
Neelakantan et al. (2015)	64
ETMo (Ours)	**67**
ETMo + NP (Ours)	66

Table 3: Results on Word Analogy task

of words and a human judgment score of the similarity measure between the two words. Similarly, SCWS-2003 consists of 2003 pairs of words, but, given with a context.

We note that our embeddings can capture only those senses that are represented by the extracted topics, and due to the restricted number of topics extracted, they might not be able to capture all the senses for a word. However, at a specific number of topics, our model is effective in capturing various senses of words in standard word similarity datasets. We demonstrate this effect qualitatively and quantitatively in this section.

For each of the datasets, we report the Spearman correlation between the human judgment score and model's similarity score computed between two words w and w'. We follow Reisinger and Mooney (2010) to compute the following similarity measures. For a pair of words w and w' and given their respective contexts c and c', we represent the cosine distance between the embeddings $E_{w,i}$ and $E_{w',j}$ as $d(E_{w,i}, E_{w',j})$.

$$
\begin{aligned}
globalSim &= d(v_g(w),) \\
avgSim &= \frac{1}{N_1 * N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} d(E_{w,i}, E_{w',j}) \\
avgSimC &= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} p(z_i|c) * p(z_j|c') * \\
&\quad d(E_{w,i}, E_{w',j})
\end{aligned}
\tag{5}
$$

N_1 and N_2 are chosen such that they satisfy $p(w_t|z_i) > p_{thres}$. $v_g(w)$ represents the *output* vector for word w, as mentioned in section 3.2. We infer the probabilities, $p(z_i|c)$ & $p(z_j|c')$ using our pre-trained topic model.

In contrast to our model, methods such as ELMo, BERT requires a document context to compute an embedding, which makes it unfair to compare on avgSim metric since it doesn't take any context into account. Additionally, ELMo gives a set of 3 different embeddings making it unclear to compare

on the avgSimC metric as well.

5.1 Quantitative Results

We present the results of our approach in Tables 1 and 2. A higher Spearman's correlation translates to a better model.

As can be seen in Table 1, our non-parametric approach clearly outperforms other multi-sense embeddings models using the *avgSim* metric on WS-353. Further, though our method focusses on sense specific embeddings and not on the global word embeddings, for the purpose of completeness, we also report our results on the *globalSim* metric. Using *globalSim*, expectedly we obtain slightly lower results since *globalSim* is more suited for global word embeddings.

In Table 2, we compare our models on the SCWS dataset. Using *avgSim* metric, our model obtain state-of-the-art results, outperforming other embeddings model. Using the *avgSimC* metric, we produce competitive results and perform better than most of the models, including Nguyen et al. (2017) which also uses topic models.

These superior results indicate the usefulness of our method to accurately capture word representations that can take into account different word senses. Additionally, our non-parametric approach consistently outperforms our baseline ETMo approach, validating our hypothesis to threshold the topics.

We also evaluate our model on the word analogy task (Mikolov et al., 2013a). [1] Our answer is correct if this word matches the correct word given in the dataset. As can seen in Table 3, our ETMo approach obtains similar results as the baseline word2vec model, and we beat other implementations.

5.2 Qualitative Comparison

We show a qualitative comparison of some polysemous words in Table 4, with the nearest neighbors of words in the table, for Glove embeddings and the embeddings trained from our model. For each of the words in Table 4, we can clearly see that the different senses of words are being effectively captured by our model whereas Glove embeddings could only capture most frequently used meaning for the word. Moreover, each of these senses can be easily correlated with the topic that these embeddings correspond to which can be seen from Table 5. Consider the word *Play*. The first sense for *play* corresponds to *Music* (topic 2). The second embedding corresponds to *Sports* (topic 7).

An interesting qualitative result is shown for the word *Network*. The nearest neighbors to Glove embeddings show that they are only able to capture one meaning which is in the subject of *Television Network*. However, our model is able to capture 3 different meanings for the word quite powerfully. The first one, which corresponds to topic 2, occurs in the context of *Television Network* which is the sense Glove was able to capture. The second sense, which corresponds to topic 5, occurs in the context of *Computer Networks*. The third sense, which corresponds to topic 6, remarkably relates to the context *Geography*.

5.3 Number of Topics Analysis

In this section, we perform a study on choosing the right number of topics(k) in Table 6. Here, topic uniqueness refers to the proportion of unique words in a topic, computed over the top words in the vocabulary. Higher the topic uniqueness score, more distinct are the obtained topics. We compute the Spearman correlation on the *avgSim* metric using the word pairs from RG-65 (Rubenstein and Goodenough, 1965). With k = 10, we obtained a topic uniqueness of 32.23, which dropped to 27.12 for k=20 topics. Thus increasing the number of topics increases overlap which harms our model as the topic weight gets divided while training the embeddings. This effect can be clearly seen in the correlation coefficient which drops from 68.5 to 66.9 for 10 & 20 topics respectively. Using k=5 improved the topic uniqueness score to 34.05, but the perplexity score (Blei et al., 2003) reduced, indicating that the topic model requires more degrees of freedom to fit the corpus. We also observed not very distinct topics at k=5 (i.e. a topic could be mixture of sports and history), resulting in reduced correlation coefficient of 67.1.

5.4 Threshold Parameter Analysis

In this section, we study the effect of p_{thres} on the model performance. We tune its value by comparing the Spearman correlation on the *avgSim* metric using the word pairs from RG-65 (Rubenstein and Goodenough, 1965). However, we hypothesize that the threshold parameter depends only on the output

[1] The word analogy task aims to answer the question of the form: *a is to b as c is to ?*. To answer the question, we compute the word vector nearest to '$v_g(b) - v_g(a) + v_g(c)$', where $v_g(w)$ represents the *output* vector for word w, as mentioned in section 3.2.

Word	Topic #	Nearest Neighbors
play	Glove	playing, played, plays, game, players, player, match, matches, games
	2	played. performance, musical, performed, plays, stage, release, song, work, time
	7	season, players, played, one, game, first, football, teams, last, year, clubs
rock	Glove	band, punk, pop, bands, album, rocks, music, indie, singer, albums, songs, rockers
	2	metal, pop, punk, members, jazz, alternative, indie, folk, band, hard, recorded, blues
	6	island, point, valley, hill, large, creek, granite, railroad, river, lake
bank	Glove	banks, banking, central, credit, bankers, financial, investment, lending, citibank
	6	river, tributary, flows, valley, side, banks, mississippi, south, north, mouth, branch
	8	company, established, central, first, group, one, investment, organisation, development
plant	Glove	plants, factory, facility, flowering, produce, reactor, factories, production
	1	plants, bird, genus, frog, rodent, flowering, fish, species, tree, endemic, asteraceae
	5	design, plants, modern, power, process, technology, standard, substance, production
war	Glove	wars, conflict, battle, civil, military, invasion, forces, fought, fighting, wartime
	4	combat, first, world, army, served, american, battle, civil, outbreak, forces
	7	series, championship, cup, fifa, champion, chess, records, wrestling, championships
network	Glove	cable, channel, television, broadcast, internet, stations, programming, radio
	2	series, program, shows, bbc, broadcast, station, channel, aired, nbc, radio, episode
	5	data, information, computer, system, applications, technology, control, standard, design
	6	light, station, car, stations, railway, commuter, lines, rail, trains, commute

Table 4: Nearest neighbours of some polysemous words for Glove, and for each sense identified by our algorithm, based on the cosine similarility. We take only those senses corresponding to topics where $p(w_t|j) > p_{thres}$.

TOPIC #	TOPIC KEYS	TOPIC NAME
1	species south india island north found small indian region family district water large east long spanish central village west area	Agriculture
2	film music album released band series show song time television single songs live rock records video release appeared episode films	Music/Television
3	party government state states united president law member general court house election served political elected national born council	Politics
4	war air army force british battle service aircraft japanese forces world military time ship fire navy command attack september car	Military
5	system formula number time called form data systems process high energy type common set space based power similar standard	Technology
6	city age county area population town located years north river south west station park line road district village income living	Geography
7	team season game league played club football games world year career player born time final cup play championship national	Sports
8	school university college company students education, public program business national research development services million service	Education
9	church book work life published century time english works art people world books language great written god early death called	Religion
10	french german france war king germany century russian part italian son chinese empire soviet republic born died emperor paris	History

Table 5: The top words for each topics according to topic modeling

# of topics	uniqueness	perplexity	$\rho \times 100$
5	34.05	9.88	67.1
10	32.23	9.78	68.5
15	29.57	9.70	67.8
20	27.12	9.65	66.9

Table 6: effect of number of topics on Spearman correlation on 50 word pairs from WordSim-353

p_{thres}	$\rho \times 100$	senses captured for *network*
1e-3	68.3	television, IT
1e-4	69.1	television, IT, transportation
1e-5	68.4	mixed senses

Table 7: Spearman's correlation $\rho \times 100$ on 50 word pairs from WS-353 and the word senses captured for *network*. The word senses are adjudged qualitatively.

of topic modeling, particularly $p(word|topic)$, and thus is independent of the this chosen subset, as can be seen in the results on other datasets. In Table 7, we can see that the optimal value for p_{thres} is 1e-4 for the non-parametric model at which it can strongly differentiate between the different senses for *network*. A higher threshold value of 1e-3 captures a fewer number of senses. A lower threshold value of 1e-5 allows training of more than the actual number of true senses leading to noisy updates, thus becoming ineffective in capturing any sense. The corresponding lower correlation coefficients in Table 7 confirm these effects quantitatively.

6 Conclusion & Future Work

In this work, we presented our approach to learn word embeddings to capture the different senses of a word. Unlike previous sense-based models, our model exploits knowledge from topic modeling to induce mixture weights in structured skip-gram approach, for learning sense specific representations. We extend this model further by pruning the embeddings conditioned on the number of word senses. Finally, we showed our model achieves state-of-the-art results on word similarity tasks, and demonstrated the strength of our model in capturing multiple word senses qualitatively. Future work should aim towards using these embeddings for downstream tasks.

References

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

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035.

Jianpeng Cheng, Zhongyuan Wang, Ji-Rong Wen, Jun Yan, and Zheng Chen. 2015. Contextual text understanding in distributional semantic space. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 133–142. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Stuart Geman and Donald Geman. 1987. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. In *Readings in computer vision*, pages 564–584. Elsevier.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 95–105.

Sujay Kumar Jauhar, Chris Dyer, and Eduard Hovy. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 683–693.

Jey Han Lau, Paul Cook, and Timothy Baldwin. 2013. unimelb: Topic modelling-based word sense induction for web snippet clustering. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 217–221.

Yang-Yin Lee, Ting-Yu Yen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. 2018. Gensense: A generalized sense retrofitting model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1662–1671.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? *arXiv preprint arXiv:1506.01070*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2015a. Learning context-sensitive word embeddings with neural tensor skip-gram model. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015b. Topical word embeddings. In *AAAI*, pages 2418–2424.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.

Massimiliano Mancini, Jose Camacho-Collados, Ignacio Iacobacci, and Roberto Navigli. 2016. Embedding words and senses together via joint knowledge-enhanced training. *arXiv preprint arXiv:1612.02703*.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Http://mallet.cs.umass.edu.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2015. Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654*.

Dai Quoc Nguyen, Dat Quoc Nguyen, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2017. A mixture model for learning multi-sense word embeddings. *arXiv preprint arXiv:1706.05111*.

Maria Pelevina, Nikolay Arefyev, Chris Biemann, and Alexander Panchenko. 2017. Making sense of word embeddings. *arXiv preprint arXiv:1708.03390*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Mohammad Taher Pilehvar and Nigel Collier. 2016. De-conflated semantic representations. *arXiv preprint arXiv:1608.01961*.

Joseph Reisinger and Raymond J Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. *arXiv preprint arXiv:1507.01127*.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

C. Shaoul and Edmonton AB: University of Alberta Westbury, C. (2010) The Westbury Lab Wikipedia Corpus. 2010. The westbury lab wikipedia dataset. Data downloaded from http://www.psych.ualberta.ca/ westburylab/downloads/westburylab.wikicorp.download.html.

Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. 2014. A probabilistic model for learning multi-prototype word embeddings. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 151–160.

Zhaohui Wu and C Lee Giles. 2015. Sense-aaware semantic analysis: A multi-prototype word representation model using wikipedia. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Heng Zhang and Guoqiang Zhong. 2016. Improving short text classification by learning vector representations of both words and hidden topics. *Knowledge-Based Systems*, 102:76–86.