# A DIFFERENTIABLE SELF-DISAMBIGUATED SENSE EMBEDDING MODEL VIA SCALED GUMBEL SOFTMAX

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We present a differentiable multi-prototype word representation model that disentangles senses of polysemous words and produces meaningful sense-specific embeddings without external resources. It jointly learns how to *disambiguate* senses given local context and how to *represent* senses using hard attention. Unlike previous multi-prototype models, our model approximates discrete sense selection in a differentiable manner via a modified Gumbel softmax. We also propose a novel human evaluation task that quantitatively measures (1) how meaningful the learned sense groups are to humans and (2) how well the model is able to disambiguate senses given a context sentence. Our model outperforms competing approaches on both human evaluations and multiple word similarity tasks.

## 1 SENSE-SPECIFIC EMBEDDING

Machine learning models for natural language processing applications often represent words with *embeddings* that are real-valued vectors. Popular word embedding models such as Word2Vec (Mikolov et al., 2013a;b) and GloVe (Pennington et al., 2014) have been instrumental in achieving state-of-the-art results on NLP tasks such as sentiment analysis (Kim, 2014; Tai et al., 2015) and textual entailment (Chen et al., 2017).

However, for *polysemous* words (those with multiple senses), learning a single vector for each word type conflates different meanings (e.g., "A hydrogen bond exists between water molecules." vs. "Do you want to buy this bond?"). This is not a new problem—Schütze (1998) demonstrates the deficiency of assigning just one vector per word—but it is more pernicious in modern models, as conflated senses can pull semantically unrelated words toward each other in the embedding space (Neelakantan et al., 2014; Pilehvar & Collier, 2016; Camacho-Collados & Pilehvar, 2018). To disentangle distinct senses in word embeddings and learn finer-grained semantic clusters, *multi-prototype word embedding* models learn multiple sense-specific embeddings for a single word (Section 7).

But what makes a good multisense word embedding? While word similarity is the most common evaluation, it has many detractors (Faruqui et al., 2016; Gladkova & Drozd, 2016): similarity is subjective and is hard to be differentiate from word relatedness. Moreover, word similarity tasks—with the exception of Stanford Contextual Word Similarity (Huang et al., 2012, SCWS)—ignore polysemous cases or are tied to specific sense inventories (Boyd-Graber et al., 2006).

Moreover, these evaluations ignore a key component of learning sense inventories: do they make sense to a human? Previous multisense embedding papers present nearest neighbors to claim their representations are interpretable and useful. Like topic models, these claims need to be rigorously verified. In Section 6, we adapt techniqes for evaluating topic models (Chang et al., 2009) to measure whether learned sense groups are internally coherent and whether humans can consistently match a learned sense vector to a word in context. Just like topic models, word embedding models that win conventional evaluations do not always make sense to humans.

We present a simple, differentiable word sense embedding model that is interpretable (measured by human evaluations) while scoring well on traditional word similarity evaluations. Our model extends the Skip-Gram Word2Vec model and simultaneously learns (1) automatic sense induction given local context and (2) sense-specific embeddings. To learn disentangled sense representations (i.e., avoid sense mixing), we approximate hard attention and preserve differentiability via a scaled variant of the Gumbel Softmax function (Section 3.2). Both qualitative and quantitative analysis show that the

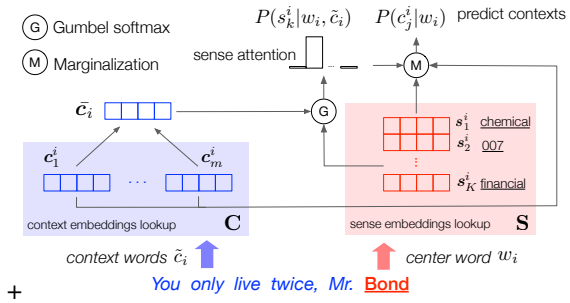


Figure 1: Network struture with an example of our GASI model which learns a set of global context embeddings **C** and a set of sense embeddings **S**

proposed variant – *Scaled Gumbel Softmax* – is critical to disambiguate senses. Figure 1 gives an overview of the structure of our model.

We compare our proposed Gumbel-Attention Sense Induction (GASI) model with previous state-of-the-art sense-specific embedding models on both word similarity tasks (Section 5) and our new crowdsourcing evaluations (Section 6). Our model performs the best on both human tasks and multiple word similarity tasks, including the SCWS task which is tailored for polysemous cases. It is also comparable to previous state-of-the-art results on other similarity tasks.

## 2 Foundations: Skip-Gram and Gumbel Softmax

Our model extends Skip-Gram Word2Vec Mikolov et al. (2013a;b). The Skip-Gram Word2Vec jointly learns word embeddings $\mathbf{W} \in \mathbb{R}^{|V|\times d}$ and context embeddings $\mathbf{C} \in \mathbb{R}^{|V|\times d}$, where $V$ is the vocabulary and $d$ is the embedding dimension, by maximizing the likelihood of the context words $c_j^i$ that surrounds a given center word $w_i$ in a context window $\tilde{c}_i$,

$$J(\mathbf{W},\mathbf{C}) \propto \sum_{w_i \in V} \sum_{c_j^i \in \tilde{c}_i} \log P(c_j^i \mid w_i; \mathbf{W}, \mathbf{C}). \tag{1}$$

Where $P(c_j^i \mid w_i)$ is estimated by a softmax over all possible context words, i.e, the whole vocabulary,

$$P(c_j^i \mid w_i; \mathbf{W}, \mathbf{C}) = \frac{\exp\left({\boldsymbol{c}_j^i}^\top \boldsymbol{w}_i\right)}{\sum_{c \in V} \exp\left(\boldsymbol{c}^\top \boldsymbol{w}_i\right)}. \tag{2}$$

In practice, $\log P(c_j^i \mid w_i)$ is approximated by negative sampling to reduce computational cost.

### 2.1 Gumbel Softmax

The Gumbel softmax Jang et al. (2016); Maddison et al. (2016) approximates the sampling of discrete random variables. Given a discrete random variable $X$ with $P(X = k) \propto \alpha_k$, $\alpha_k \in (0, \infty)$, the Gumbel-max Gumbel & Lieblein (1954); Maddison et al. (2014) refactors the sampling of $X$ into

$$X = \arg\max_k (\log \alpha_k + g_k), \tag{3}$$

where the Gumbel noise $g_k = -\log(-\log(u_k))$ and $u_k$ are i.i.d samples drawn from Uniform(0, 1).

The Gumbel softmax approximates the sampling results $\text{one_hot}(\arg\max_k(\log \alpha_k + g_k))$ by,

$$y_k = \text{softmax}((\log \alpha_k + g_k)/\tau). \tag{4}$$

## 3 Gumbel-Attention Sense Induction (GASI)

Building on these foundations, we now introduce our model (GASI, and along the way introduce a soft-attention stepping-stone, SASI); afterward, we will compare the model on both traditional

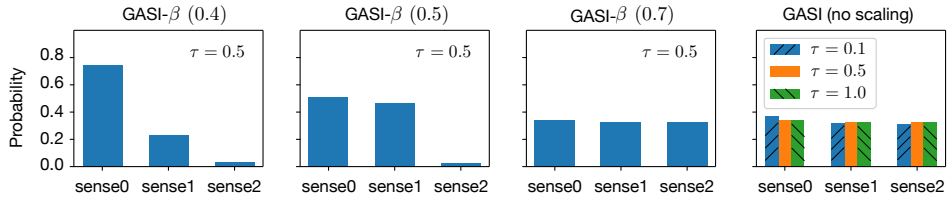


Figure 2: As the scale factor $\beta$ increases, the sense selection distribution for "bond" given examples from SemCor3.0 for synset "bond.n.02" becomes flatter, indicating less disambiguated sense vectors.

evaluation metrics and interpretability. The critical component of our model is that we model the sense selection probability, which can be interpreted as sense attention over contexts, into the Skip-Gram model while preserving the original objective through marginalization (Figure 1). By using Gumbel Softmax, our model both approximates discrete sense selection and is differentiable. Previous models are either non-differentiable or otherwise complicate inference through hard attention with reinforcement learning methods (Lee & Chen, 2017).

### 3.1 Attentional Sense Induction Framework

**Embedding Parameters** We learn a context embedding matrix $\mathbf{C} \in \mathbb{R}^{|V| \times d}$ and a sense embedding tensor $\mathbf{S} \in \mathbb{R}^{|V| \times K \times d}$. Unlike previous work (Neelakantan et al., 2014; Lee & Chen, 2017), no extra embeddings are kept for sense induction.

**Number of Senses** For simplicity and consistency with most of previous work, we present our model with a fixed number of senses $K$.[1] We can prune duplicate senses by a model-specific pruning threshold $\lambda$ estimated from learned embeddings,

$$\lambda = \frac{1}{2}(\text{mean}(D_{dup}) + \text{mean}(D_{nn})), \tag{5}$$

where $D_{dup}$ is a set of cosine distances between duplicate senses (senses of the same word that are nearest neighbors of each other) of sampled words and $D_{nn}$ is a set of cosine distances between different nearest-neighbor words.[2]

**Sense Attention in Objective Function** Assuming a center word $w_i$ has senses $\{s_1^i, s_2^i, \ldots, s_K^i\}$, the original Skip-Gram likelihood given local context $\tilde{c}_i$ can be written as marginal distribution over all senses of $w_i$ with the sense induction probability $P(s_k^i \mid w_i, \tilde{c}_i)$.

$$P(c_j^i \mid w_i, \tilde{c}_i) = \sum_{k=1}^{K} P(c_j^i \mid s_k^i) \underbrace{P(s_k^i \mid w_i, \tilde{c}_i)}_{\text{attention}}, \tag{6}$$

Replacing $P(c_j^i \mid w_i)$ in Equation 1 with Equation 6 gives our objective function,

$$J(\mathbf{S}, \mathbf{C}) \propto \sum_{w_i \in V} \sum_{c_j^i \in \tilde{c}_i} \log \sum_{k=1}^{K} P(c_j^i \mid s_k^i) P(s_k^i \mid w_i, \tilde{c}_i). \tag{7}$$

**Lower Bound the Objective for Negative Sampling** Like the Skip-Gram objective (Equation 2), we model the likelihood of a context word given the center sense $P(c_j^i \mid s_k^i)$ using softmax,

$$P(c_j^i \mid s_k^i) = \frac{\exp\left(\boldsymbol{c}_j^{i\top} \boldsymbol{s}_k^i\right)}{\sum_{j=1}^{|V|} \exp\left(\boldsymbol{c}_j^\top \boldsymbol{s}_k^i\right)}, \tag{8}$$

where the bold symbol $\boldsymbol{s}_k^i$ looks up the embedding of sense $s_k^j$ from $\mathbf{S}$, and $\boldsymbol{c}_j$ looks up the context embedding of word $c_j$ from $\mathbf{C}$.

Computing the softmax over the vocabulary is time-consuming. We want to adopt negative sampling to approximate $\log P(c_j^i \mid s_k^i)$, which does not exist explicitly in our objective function (Equation 7).[3]

[1] We can set different number of senses based on word frequency in the training, details in Appendix B.3.

[2] We elaborate the details and discuss our choices in Appendix B.

[3] Deriving the negative sampling requires the logarithm of a softmax Goldberg & Levy (2014).

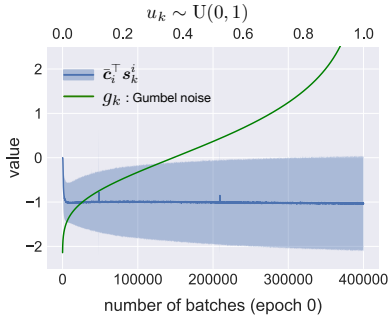


Figure 3: Our hard attention mechanism is approximated with Gumbel softmax on the context-sense dot product $\bar{\boldsymbol{c}}_i^\top \boldsymbol{s}_k^i$ (Equation 14), whose mean and std plotted here as a function of iteration. The shadowed area shows that it has a smaller scale than the gumbel noise $g_k$, such that $g_k$, rather than the embeddings, dominates the sense attention.

However, given the concavity of the logarithm function, we can apply Jensen's inequality,

$$\log \sum_{k=1}^{K} P(c_j^i \mid s_k^i) P(s_k^i \mid w_i, \tilde{c}_i) \geq \sum_{k=1}^{K} P(s_k^i \mid w_i, \tilde{c}_i) \log P(c_j^i \mid s_k^i), \tag{9}$$

and create a lower bound of the objective. Maximizing this lower bound gives us a *tractable objective*,

$$J(\mathbf{S}, \mathbf{C}) \propto \sum_{w_i \in V} \sum_{c_j^i \in \tilde{c}_i} \sum_{k=1}^{K} P(s_k^i \mid w_i, \tilde{c}_i) \log P(c_j^i \mid s_k^i), \tag{10}$$

where $\log P(c_j^i \mid s_k^i)$ is estimated by negative sampling Mikolov et al. (2013b),

$$\log \sigma({\boldsymbol{c}_j^i}^\top \boldsymbol{s}_k^i) + \sum_{j=1}^{n} \mathbb{E}_{c_j \sim P_n(c)}[\log \sigma(-\boldsymbol{c}_j^\top \boldsymbol{s}_k^j))], \tag{11}$$

**Modeling Sense Attention** We can model the attention term, *contextual sense induction distribution*, with soft attention; we call the resulting model soft-attention sense induction (SASI); although it is a stepping stone to our final model, we compare against it in our experiments as it helps isolate the contributions of hard attention. In SASI, the sense attention is conditioned on the entire local context $\tilde{c}_i$ with softmax:

$$P(s_k^i \mid w_i, \tilde{c}_i) = \frac{\exp\left(\bar{\boldsymbol{c}}_i^\top \boldsymbol{s}_k^i\right)}{\sum_{k=1}^{K} \exp\left(\bar{\boldsymbol{c}}_i^\top \boldsymbol{s}_k^i\right)}, \tag{12}$$

where $\bar{\boldsymbol{c}}_i$ is the mean of the context vectors in $\tilde{c}_i$.

### 3.2 Scaled Gumbel Softmax for Sense Disambiguation

To reduce separate senses and learn *distinguishable sense representations*, we implement *hard* attention in our full model, GASI. To preserve differentiability and circumvent the difficulties in training reinforcement learning (Sutton & Barto, 1998), we apply the reparameterization trick with Gumbel softmax (Section 2.1) to our sense attention function (Equation 12) and make a continuous relaxation.

**Vanilla Gumbel Attention** The discrete sense sampling from Equation 12 can be refactored by

$$\boldsymbol{z}^{\boldsymbol{i}} = \text{one_hot}(\arg\max_k({\bar{\boldsymbol{c}}_i}^\top \boldsymbol{s}_k^i + g_k)), \tag{13}$$

and the hard attention is approximated with

$$y_k^i = \text{softmax}(({\bar{\boldsymbol{c}}_i}^\top \boldsymbol{s}_k^i + g_k)/\tau). \tag{14}$$

**Scaled Gumbel Softmax for Sense Disambiguation** Gumbel softmax learns a flat distribution over senses even with low temperatures (Figure 2): the dot product $\bar{\boldsymbol{c}}_i^\top \boldsymbol{s}_k^i$ is too small compared to the Gumbel noise $g_k$ (Figure 3).[4] Thus we use a scaling factor $\beta$ to reduce the randomness,[5] and tune

[4]Float32 precision, the saturation of $\log(\sigma(\cdot))$ and gradient vanishing result in a small range of $\bar{\boldsymbol{c}}_i^\top \boldsymbol{s}_k^i$.
[5]Normalizing $\bar{\boldsymbol{c}}_i^\top \boldsymbol{s}_k^i$ or directly using $\log P(s_k^i \mid w_i, \tilde{c}_i)$ results in a similar outcome.

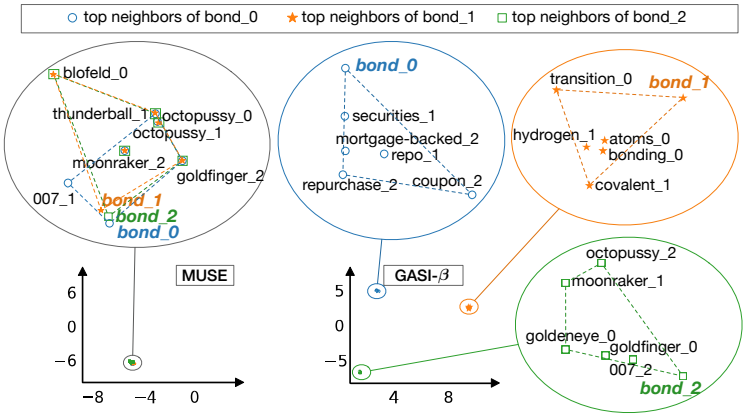


Figure 4: t-SNE projections of nearest neighbors for "bond" by *hard-attention* models: 1) previous SOTA model MUSE (RL-based); 2) our proposed GASI-$\beta$. Trained on same dataset and vocabulary, both models learn three vectors per word. Here, $word_i$ represent the $i$-th vector for $word$. Our GASI (right) learns three distinct senses of "bond" while MUSE (left) learns overlapping senses.

it as a hyperparameter.[6]

$$\gamma_k^i = \text{softmax}((\bar{\boldsymbol{c}}_i^\top \boldsymbol{s}_{ik} + \beta g_k)/\tau), \tag{15}$$

We use GASI-$\beta$ to identify the GASI model with scaling factor. This modification is critical for learning *distinguishable senses* (Figure 2, Table 1, and Table 4).

**Final Objective Function** The objective function of our GASI-$\beta$ model is

$$J(\mathbf{S}, \mathbf{C}) \propto \sum_{w_i \in V} \sum_{w_c \in c_i} \sum_{k=1}^{K} \text{softmax}((\bar{\boldsymbol{c}}_i^\top \boldsymbol{s}_{ik} + \beta g_k)/\tau) \log P(w_c \mid s_k^i). \tag{16}$$

## 4 Training Settings

For fair comparisons, we try to remain consistent with previous work (Huang et al., 2012; Neelakantan et al., 2014; Lee & Chen, 2017) in all aspects of training. In particular, we train GASI on the same April 2010 Wikipedia snapshot (Shaoul C., 2010) with 1B tokens the same vocabulary released by Neelakantan et al. (2014); set the number of senses $K = 3$ and dimension $d = 300$ for each word unless otherwise specified. More details are in Appendix A. We fix the temperature $\tau = 0.5$,[7] and tune the scaling factor $\beta$ from $\{0.1, 0.2, ..., 0.9\}$ on the AvgSimC measure for the contextual word similarity task (Section 5). The optimal scaling factor $\beta$ is 0.4.

If not reprinted, the numbers in this paper for competing models are either computed with pre-trained embeddings released by authors or trained on released code.[8]

## 5 Word Similarity Evaluation

We first compare our GASI and GASI-$\beta$ model with previous work on standard word similarity tasks before turning to interpretability experiments. Each task has word pairs with a similarity/relatedness score. For evaluation, we measure Spearman's rank correlation $\rho$ (Spearman, 1904) between word embedding similarity and the gold similarity judgements: higher scores imply the model captures semantic similarities consistent with the trusted similarity scores.

[6] Learning $\beta$ instead of fixing it as a hyperparameter does not successfully disambiguate senses.

[7] This is similar to the experiment settings for Gumbel softmax in Maddison et al. (2016)

[8] We adopt the numbers for Li & Jurafsky (2015) from Lee & Chen (2017) and tune the PFT-GM (Athiwaratkun et al., 2018) model on the same 1B corpus and vocabulary as previous works using https://github.com/benathi/multisense-prob-fasttext with suggested hyperparameters and select the best results.

| Model | MaxSimC | AvgSimC |
|---|---|---|
| Huang et al. (2012)-50d | 26.1 | 65.7 |
| MSSG-6K | 57.3 | 69.3 |
| MSSG-30K | 59.3 | 69.2 |
| Tian et al. (2014) | 63.6 | 65.4 |
| Li & Jurafsky (2015) | 66.6 | 66.8 |
| Qiu et al. (2016) | 64.9 | 66.1 |
| Bartunov et al. (2016) | 53.8 | 61.2 |
| MUSE_Boltzmann | 67.9 | 68.7 |
| SASI | 55.1 | 67.8 |
| GASI(w/o scaling) | **68.2** | 68.3 |
| GASI-$\beta$ | 66.4 | **69.5** |

Table 1: Spearman's correlation $100\rho$ on SCWS. All models are trained on the 1B token data and learn 300d embeddings except for Huang et al.

**Contextual Word Similarity** Tailored for sense embedding evaluation, Stanford Contextual Word Similarities (Huang et al., 2012, SCWS) has 2003 word pairs and similarity scores with sentential context. Moreover, the word pairs and their contexts reflect homonymous and polysemous words. Therefore, we use this dataset to tune our hyperparameters.

To compute the word similarity with senses we use two metrics Reisinger & Mooney (2010) that take context and sense disambiguation into account: **MaxSimC** computes the cosine similarity $\cos(s_1^*, s_2^*)$ between the two most probable senses $s_1^*$ and $s_2^*$ that maximizes $P(s_k^i \mid w_i, \tilde{c}_i)$. **AvgSimC** weights average similarity over the combinations of all senses,

$$\sum_{i=1}^{K}\sum_{i=j}^{K} P(s_i^1 \mid w_1, \tilde{c}_1)P(s_j^2 \mid w_2, \tilde{c}_2)\cos(s_i^1 s_j^2). \tag{17}$$

We first compare variants of our model with multi-prototype sense embedding models (Table 1), including two previous state-of-the-art models: the clustering-based Multi-Sense Skip-Gram model (Neelakantan et al., 2014, MSSG) on AvgSimC metric and the RL-based Modularizing Unsupervised Sense Embeddings (Lee & Chen, 2017, MUSE) on MaxSimC. All three are better than the baseline Skip-Gram model (65.2 using the word embedding).

GASI better captures similarity than SASI, corroborating that hard attention aids word sense selection. Moreover, GASI with scaling ($\beta$) has better MaxSimC than all other models; however, it learns a flat sense distribution (Figure 2). GASI-$\beta$ has the best AvgSimC and a competitive MaxSimC. While MUSE has a higher MaxSimC than GASI-$\beta$, it fails to distinguish senses as well (Figure 4, Section 6).

The Probabilistic FastText Gaussian Mixture (Athiwaratkun et al., 2018, PFT-GM) achieves state-of-the-art results on multiple non-contextual word similarity tasks (Table 2). Since it does not estimate the sense induction distribution based on local context, we compute the correlation score (66.4) for PFT-GM with MaxSim (Equation 18). Our GASI-$\beta$ is comparable to it on MaxSim with the same 66.4, and has better correlation on AvgSimC (69.5).

**Non-Contextual Word Similarity** We also evaluate our model on the non-contextual word similarity datasets: RG-65 (Rubenstein & Goodenough, 1965); SimLex-999 (Hill et al., 2015); WS-353 (Finkelstein et al., 2002); MEN-3k (Bruni et al., 2014); MC-30 (Miller & Charles, 1991); YP-130 (Yang & Powers, 2006); MTurk-287 (Radinsky et al., 2011); MTurk-771 (Halawi et al., 2012); RW-2k (Luong et al., 2013). Similar to Lee & Chen (2017) and Athiwaratkun et al. (2018), we compute the word similarity based on senses by **MaxSim** (Reisinger & Mooney, 2010), which maximizes the cosine similarity over the combination of all sense pairs and does not require local contexts,

$$\text{MaxSim}(w_1, w_2) = \max_{0\le i\le K, 0\le j\le K} \cos(s_i^1, s_j^2). \tag{18}$$

This metric evaluates the quality of each specific sense embedding, rather than the average embeddings. GASI-$\beta$ has better correlation on three datasets and is competitive on the rest (Table 2) and remains

| Dataset | MSSG-30K | MSSG-6K | MUSE_Boltzmann | GASI | GASI-$\beta$ | PFT-GM |
|---|---|---|---|---|---|---|
| SimLex-999 | 31.80 | 28.65 | 39.61 | 40.14 | **41.68** | 40.19 |
| WS-353 | 65.69 | 67.42 | 68.41 | 68.49 | **69.36** | 68.6 |
| MEN-3k | 65.99 | 67.10 | 74.06 | 73.13 | 72.32 | **77.40** |
| MC-30 | 67.79 | 76.02 | 81.80 | 82.47 | **85.27** | 74.63 |
| RG-65 | 73.90 | 64.97 | **81.11** | 77.19 | 79.77 | 79.75 |
| YP-130 | 40.69 | 42.68 | 43.56 | 49.82 | 56.34 | **59.39** |
| MT-287 | 65.47 | 64.04 | 67.22 | 67.37 | 66.13 | **69.66** |
| MT-771 | 61.26 | 58.83 | 64.00 | 66.65 | 66.70 | **68.91** |
| RW-2k | 42.87 | 39.24 | **48.46** | 47.22 | 47.69 | 45.69 |

Table 2: Spearman's correlation $100\rho$ on non-contextual word similarity measured by MaxSim. GASI-$\beta$ outperforms the other models on three datasets are competitive on the others. Note that PFT-GM are trained with two components/senses while other models learn three senses.

competitive without scaling. GASI performs better than MUSE, the other hard-attention multi-prototype model, on six datasets and worse on three. Taken as a whole, our results on these non-contextual word similarity tasks indicate that we have not lowered the quality of our word representations in the process of introducing sense-specific mechanisms.

## 6 CROWDSOURCING EVALUATION

GASI can capture word similarity (Section 5), but do the learned representations make sense? Could a human use them to help build a dictionary? If you show a human the senses, can they understand why a model would assign a sense to that context? In this section we evaluate whether the representations make sense to human consumers of multisense models.

**Qualitive analysis** Previous papers use nearest neighbors of a few examples to qualitatively argue that their models have captured meaningful senses of words. We also give an example in Figure 4, which provides an intuitive view on how the learned senses are clustered by visualizing the nearest neighbors of word "bond" using t-SNE projection (Maaten & Hinton, 2008). Our proposed model (right) disentangles the three sense of "bond" clearly and learns three distinct sense vectors.

However, the examples can be cherry-picked and lack standards. This problem also bedeveled topic modeling until the introduction of rigorous human evaluation (Chang et al., 2009). We adapt both aspects Chang *et al*'s evaluations: *word intrusion* (Schnabel et al., 2015) to evaluate whether individual senses are coherent and *topic intrusion*—rather sense intrusion in this case—to evaluate whether humans agree with models' sense assignments *in context*. Both crowdsourcing tasks collect human inputs on Figure-Eight.

We compare our models with the two previous state-of-the-art multi-prototype sense embeddings models that disambiguate senses given local context, i.e., MSSG (Neelakantan et al., 2014) and MUSE (Lee & Chen, 2017).[9]

### 6.1 WORD INTRUSION FOR SENSE COHERENCE

Schnabel et al. (2015) suggests a "good" word embedding should have coherent neighbors and evaluate coherence by *word intrusion* they present Turkers a group of four words, three of which are close neighbors while one of which is an "intruder", and ask the Turkers to find the intrusion word. If the original sense/topic/embedding makes sense, contributors will easily spot the word that "does not belong".

Similarly, we examine the coherency of the ten nearest neighbors used in the *contextual word sense selection* task (Section 6.2) by randomly replace one neighbor with an "intruder", and ask the contributor to find the intrusion word Figure 5. We generate three questions with different intruders for each sense and collect three judgements per question. We consider the "intruder" to be corretly selected if at least two judgements are correct.

[9]MSSG has two settings, we run human evaluation with MSSG-30K which achieves higher correlation score with MaxSimC on SCWS, indicating better quality of each sense vector given local context.

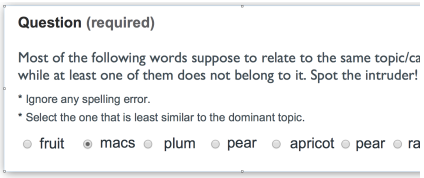


Figure 5: Word intrusion task prompt

| Model | Sense-level Accuracy | Judgement-level Accuracy | Aggrement |
|---|---|---|---|
| MUSE | 67.33 | 62.89 | 0.73 |
| MSSG-30K | 69.33 | 66.67 | 0.76 |
| GASI-$\beta$ | **71.33** | **67.33** | **0.77** |

Table 3: Word intrusion evalutations on top ten nearest neighbors of sense embeddings.

**Generate Intruder** Like Chang et al. (2009), we want the "intruder" to not be too different in terms of frequency to the target set but not too similar semantically. For sense $s_i^m$ of word type $w_i$, we randomly select a word from the neighbors of another sense $s_i^n$ of $w_i$ but with a low threshold, i.e., any words that has cosine similarity larger than 0.0 can be viewed as a neighbor. We discuss further refinement of intruder selection in the discussion section.

**Result and Analysis** All three models have comparable model accuracy. GASI-$\beta$ learns senses that have the highest coherency among top ten nearest neighbors while MUSE learns more sense mixtures.

**Agreement** We use the aggregated confidence score provided by Figure-Figure to estimate the level of agreement between multiple contributors.[10] The agreements are adequately high for all models and our GASI-$\beta$ achieves the highest agreement, indicates that the senses represented by nearest neighbors learned by GASI-$\beta$ are easier to interpret for human.

### 6.2 CONTEXTUAL WORD SENSE SELECTION

The previous task measures whether individual senses are coherent. In this task, we measure whether the learned senses by sense embedding models make sense human and evaluate the models' ability in disambiguate senses in context.

**Task Description** Given a target word in context, we ask a contributor to select which sense group learned by a model that best fits the sentence. Each sense group is described by its top ten distinct nearest neighbors (Figure 6).[11]

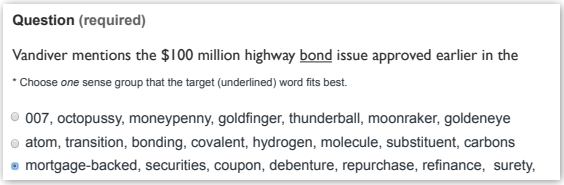


Figure 6: An example (target: *bond*) of the *contextual word sense selection* task; each option contains top ten nearest neighbors of a sense embedding learned by the model; senses in this example are from our GASI-$\beta$ (1. 007; 2. chemical; 3. financial).

**Data Collection** We select fifty nouns with five sentences from SemCor 3.0 (Miller et al., 1994). We first filter all word senses in the dataset that have fewer than ten sentences, then select the top fifty nouns among the remaining via the number of synsets in WordNet (Miller & Fellbaum, 1998). For each noun, we randomly select five sentences, each as one question in this task.

**Metrics** For each model, we collect three judgements for each question. For each question, if at least two contributors select the same sense as the model does, we consider the model's selection is consistent with that of humans. We also compute the sense induction probability $P$ assigned to the human choices by the model, indicating the model's confidence in sense selection. $P = 1/3$ indicates the model learns flat sense induction distribution is unable to disambiguate senses.

**Sense disambiguation and interpretability** If a user can consistently pick the same sense as the model, it means that three things are true. 1) human is able to interpret each group of nearest neighbor words (as measured by the previous experiment); 2) the senses are distinguishable enough from each other for the human to make one selection; 3) the human's selection is consistent with the model's.

[10] https://success.figure-eight.com/hc/en-us/articles/201855939-How-to-Calculate-a-Confidence-Score

[11] We shuffle the choices for questions with the same target word.

| Model | Accuracy | $P$ | Agreement |
|---|---|---|---|
| MUSE | 28.0 | 0.33 | 0.68 |
| MSSG-30K | 44.5 | 0.37 | 0.73 |
| GASI (w/o scaling) | 33.8 | 0.33 | 0.68 |
| GASI-$\beta$ | **50.0** | **0.48** | **0.75** |

Table 4: Human-model consistency on *contextual word sense selection*, where $P$ is the average probability assigned by the model to the human choices. GASI-$\beta$ is more consistent with human than baseline models.

**Results and Analysis** GASI-$\beta$ selects senses that are most consistent with humans; it has the highest accuracy and assigns the largest probability assigned to the human choices (Table 4, left). Thus, GASI-$\beta$ produces sense embeddings that are both more interpretable and distinguishable, and has stronger ability in disambiguating sense.[12] The canonical GASI model without scaling factor, however, has low consistency and flat sense distribution. Therefore, our proposed modification to Gumbel Softmax is critical in learning sense disambiguation.

**Agreement** We use the confidence score computed by Figure-Eight to estimate the rater's agreement for this task as well. Our GASI-$\beta$ achieves the hight agreement while both MUSE and GASI without scaling have the lowest.

### 6.3 Word Similarity vs. Sense Disambiguation

The evaluation results on word similarity tasks (Section 5) and human evaluations (Section 6) are inconsistent for several models. Among all the models that learn sense-specific embedding and support sense disambiguation given local context, our GASI, GASI-$\beta$ model and the MUSE model are equally competitive in overall word similarity evaluations and achieve at least close to state-of-the-art results (Table 1 and Table 2). However, both GASI and MUSE perform poorly in the human evaluation task while GASI-$\beta$ performs the best (Table 4). Both canonical GASI model and MUSE fail to learn distinguishable sense embeddings and cannot actually disambiguate senses given local context. Therefore, achieving high word similarity evaluation performance does not necessarily indicate "good" sense embeddings quality. Our proposed human evaluation task – *contextual word sense selection* – provides complementary evaluation.

## 7 Related Work

Schütze (1998) introduces context-group discrimination for senses and uses the centroid of context vectors as a sense representation. Other work induces senses by context clustering (Purandare & Pedersen, 2004) or probabilistic mixture models (Brody & Lapata, 2009). Reisinger & Mooney (2010) first introduce multiple sense-specific vectors for each word, inspiring other multi-prototype sense embedding models.

Generally, to address polysemy in word embeddings, some previous work trained on annotated sense corpora (Iacobacci et al., 2015) or external sense inventories (Labutov & Lipson, 2013; Chen et al., 2014; Jauhar et al., 2015; Chen et al., 2015; Wu & Giles, 2015; Pelevina et al., 2016); Rothe & Schütze (2015; 2017) extend word embeddings to lexical resources without training; others induce senses via multilingual parallel corpora (Guo et al., 2014; Šuster et al., 2016; Ettinger et al., 2016).

We mainly contrast our GASI to unsupervised monolingual multi-prototype models along two dimensions: *sense induction methodology* and *differentiability*. On the first dimension, Huang et al. (2012) and Neelakantan et al. (2014) induce senses by context clustering; Tian et al. (2014) model a corpus-level sense distribution; Li & Jurafsky (2015) model the sense assignment as a Chinese Restaurant Process; Qiu et al. (2016) induce senses by minimizing an energy function on a context-depend network; Bartunov et al. (2016) model the sense assignment as a steak-breaking process; Nguyen et al. (2017) model the sense embeddings as a weighted combination of topic vectors with pre-computed

[12] Workers tends to select the first choice when the choices are duplicated, which explains why MUSE achieves lower accuracy than random choice.

weights by topic models; Athiwaratkun & Wilson (2017) and Athiwaratkun et al. (2018) model word representations as Gaussian Mixture embeddings where each Gaussian component captures different senses; Lee & Chen (2017) computes sense distribution by a separate set of sense induction vectors; while our GASI introduces sense attention by marginalizing the likelihood of contexts over senses and induces senses by local context vectors; the most similar structure actually is a bilingual model (Šuster et al., 2016) except it does not introduce lower bound for negative sampling. On the second dimension, for models that learn sense disambiguation given local context, most of which are *non-differentiable* and discretely select senses. The two exceptions are: Šuster et al. (2016) use weighted vectors over senses; Lee & Chen (2017) implement hard attention with reinforcement learning to mitigate the non-differentiability. In contrast, GASI keeps full differentiability by reparameterization and approximates discrete sense sampling with scaled Gumbel softmax.

## 8 Conclusion

We present a differentiable Gumbel Attention Sense Induction (GASI) model that learns both distinguishable and meaningful sense representations for words. It applies hard attention to simultaneously induce and embed word senses from unlabeled monolingual corpora, and approximates the discrete sense sampling with *Scaled Gumbel Softmax*. The proposed scaling factor is critical for our hard-attention model to disambiguate senses given local context. Due to the lack of standard intrinsic evaluation for sense-specific embeddings, we propose a novel crowdsourcing *contextual word sense selection* task to quantitatively evaluate the semantic meaningfulness of sense embeddings. It measures how consistent the model's sense selections are to that of humans given context sentence. We further evaluate the coherency of the learned senses represented each by its nearest neighbors through the *word intrusion* human evaluation task. The scaled version GASI-$\beta$ of our sense embedding model achieves higher accuracy on both consistent and coherent human evaluations than competing approaches while maintaining sate-of-the-art results on multiple traditional word similarity tasks, including the SCWS task which is tailored for polysemous cases. It is also comparable to previous state-of-the-art results on other similarity tasks. On the other hand, the inconsistent performance of several sense embedding models on human evaluation tasks and word similarity tasks indicate our proposed *contextual word sense selection* is complementary to traditional word similarity evaluations.

We believe that evaluating the interpretability of sense representations is important not only for building better models but for improving our understanding of language. Creating intruder words and tasks that better align with real-world tasks of lexicographers or linguists would help better ground word sense embedding evaluations.

## References

Ben Athiwaratkun and Andrew Wilson. Multimodal word distributions. In *Proceedings of the Association for Computational Linguistics*, 2017.

Ben Athiwaratkun, Andrew Wilson, and Anima Anandkumar. Probabilistic fasttext for multi-sense word embeddings. In *Proceedings of Empirical Methods in Natural Language Processing*, 2018.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. Breaking sticks and ambiguities with adaptive skip-gram. In *Proceedings of Artificial Intelligence and Statistics*, 2016.

Jordan L. Boyd-Graber, Sonya S. Nikolova, Karyn A. Moffatt, Kenrick C. Kin, Joshua Y. Lee, Lester W. Mackey, Marilyn M. Tremaine, and Maria M. Klawe. Participatory design with proxies: Developing a desktop-PDA system to support people with aphasia. In *International Conference on Human Factors in Computing Systems*, 2006.

Samuel Brody and Mirella Lapata. Bayesian word sense induction. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, 2009.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49, 2014.

Jose Camacho-Collados and Taher Pilehvar. From word to sense embeddings: A survey on vector representations of meaning. *arXiv preprint arXiv:1805.04032*, 2018.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Proceedings of Advances in Neural Information Processing Systems*, 2009.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced lstm for natural language inference. In *Proceedings of the Association for Computational Linguistics*, 2017.

Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. Improving distributed representation of word sense via wordnet gloss composition and context clustering. In *Proceedings of the Association for Computational Linguistics*, pp. 15–20. Association for Computational Linguistics, 2015.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. A unified model for word sense representation and disambiguation. In *Proceedings of Empirical Methods in Natural Language Processing*. Citeseer, 2014.

Allyson Ettinger, Philip Resnik, and Marine Carpuat. Retrofitting sense-specific word vectors using parallel text. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2016.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the Association for Computational Linguistics*, 2016.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1), 2002.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76 (5), 1971.

Anna Gladkova and Aleksandr Drozd. Intrinsic evaluations of word embeddings: What can we do better? In *RepEval@ACL*, 2016.

Yoav Goldberg and Omer Levy. word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

Emil Julius Gumbel and Julius Lieblein. Statistical theory of extreme values and some practical applications: a series of lectures. 1954.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. Learning sense-specific word embeddings by exploiting bilingual resources. In *Proceedings of International Conference on Computational Linguistics*, 2014.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. Large-scale learning of word relatedness with constraints. In *Knowledge Discovery and Data Mining*, 2012.

Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4), 2015.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the Association for Computational Linguistics*, 2012.

Ignacio Iacobacci, Taher Mohammad Pilehvar, and Roberto Navigli. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the Association for Computational Linguistics*, pp. 95–105. Association for Computational Linguistics, 2015.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Sujay Kumar Jauhar, Chris Dyer, and Eduard Hovy. Ontologically grounded multi-sense representation learning for semantic vector space models. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2015.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of Empirical Methods in Natural Language Processing*, 2014.

Igor Labutov and Hod Lipson. Re-embedding words. In *Proceedings of the Association for Computational Linguistics*, 2013.

Guang-He Lee and Yun-Nung Chen. Muse: Modularizing unsupervised sense embeddings. In *Proceedings of Empirical Methods in Natural Language Processing*, 2017.

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

Thang Luong, Richard Socher, and Christopher Manning. Better word representations with recursive neural networks for morphology. In *Conference on Computational Natural Language Learning*, 2013.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov), 2008.

Chris J Maddison, Daniel Tarlow, and Tom Minka. A* sampling. In *Proceedings of Advances in Neural Information Processing Systems*, 2014.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems*, 2013b.

George Miller and Christiane Fellbaum. *Wordnet: An electronic lexical database*. MIT Press Cambridge, 1998.

George A Miller and Walter G Charles. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6, 1991.

George A Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G Thomas. Using a semantic concordance for sense identification. In *Proceedings of the workshop on Human Language Technology*, 1994.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of Empirical Methods in Natural Language Processing*, 2014.

Dai Quoc Nguyen, Dat Quoc Nguyen, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. A mixture model for learning multi-sense word embeddings. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*, 2017.

Maria Pelevina, Nikolay Arefiev, Chris Biemann, and Alexander Panchenko. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, 2016.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of Empirical Methods in Natural Language Processing*, 2014.

Mohammad Taher Pilehvar and Nigel Collier. De-conflated semantic representations. In *Proceedings of Empirical Methods in Natural Language Processing*, 2016.

Amruta Purandare and Ted Pedersen. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Conference on Computational Natural Language Learning*, 2004.

Lin Qiu, Kewei Tu, and Yong Yu. Context-dependent sense embedding. In *Proceedings of Empirical Methods in Natural Language Processing*, 2016.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the World Wide Web Conference*, 2011.

Joseph Reisinger and Raymond J. Mooney. Multi-prototype vector-space models of word meaning. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

Sascha Rothe and Hinrich Schütze. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the Association for Computational Linguistics*, 2015.

Sascha Rothe and Hinrich Schütze. Autoextend: Combining word embeddings with semantic resources. *Computational Linguistics*, 43(3), 2017.

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

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. Evaluation methods for unsupervised word embeddings. In *Proceedings of Empirical Methods in Natural Language Processing*, pp. 298–307, 2015.

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

Westbury C Shaoul C. *The Westbury Lab Wikipedia Corpusa*. 2010.

Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15, 1904.

Simon Šuster, Ivan Titov, and Gertjan van Noord. Bilingual learning of multi-sense embeddings with discrete autoencoders. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2016.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the Association for Computational Linguistics*, 2015.

Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. A probabilistic model for learning multi-prototype word embeddings. In *Proceedings of International Conference on Computational Linguistics*, 2014.

Zhaohui Wu and C Lee Giles. Sense-aaware semantic analysis: A multi-prototype word representation model using wikipedia. In *Association for the Advancement of Artificial Intelligence*, pp. 2188–2194. Citeseer, 2015.

Dongqiang Yang and David Martin Powers. *Verb similarity on the taxonomy of WordNet*. Masaryk University, 2006.

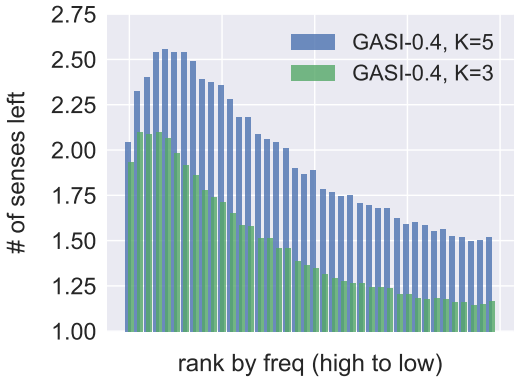


Figure 7: Histogram of number of senses left after post-training pruning for two models: GASI-0.4 initialized with three senses and GASI-0.4 initialized with five senses. We rank the number of senses of words by their frequency from high to low.

## A TRAINING DETAILS

During training, we fix the window size to five and the dimensionality of the embedding space to 300 for comparison to previous work. We initialize both sense and context embeddings randomly within U(-0.5/dim, 0.5/dim) as in Word2Vec. We set the initial learning rate to 0.01; it is decreased linearly until training concludes after 5 epochs. The batch size is 512, and we use five negative samples per center word-context pair as suggested by Mikolov et al. (2013a). We train our model on the GeForce GTX 1080 Ti, and our implementation (using pytorch 3.0) takes $\sim$ 6 hours to train one epoch on the April 2010 Wikipedia snapshot Shaoul C. (2010) with 100k vocabulary. For comparison, our implementation of Skip-Gram on the same framework takes $\sim$ 2 hours each epoch.

## B NUMBER OF SENSES

For simplicity and consistency with most of previous work, we present our model with a fixed number of senses $K$.

### B.1 POST-TRAINING PRUNING

For words that do not have multiple senses or have most senses appear very low-frequently in corpus, our model (as well as many previous models) learns duplicate senses. We can easily remove such duplicates by *pruning* the learned sense embeddings with a threshold $\lambda$. Specifically, for each word $w_i$, if the cosine distance between any of its sense embeddings $(\boldsymbol{s}^i_m, \boldsymbol{s}^i_n)$ is smaller than $\lambda$, we consider them to be duplicates. After discovering all duplicate pairs, we start pruning with the sense $s^i_k$ that has the most duplications and keep pruning with the same strategy until no more duplicates remain.

**Model-specific pruning** We estimate a model-specific threshold $\lambda$ from the learned embeddings instead of deciding it arbitrary. Therefore, this pruning methods is also applicable to other sense embedding models. We first sample 100 words from the negative sampling distribution over the vocabulary. Then, we retrieve the five nearest neighbors (from all senses of all words) to each sense of each sampled word. If one of a word's own senses appears as a nearest neighbor, we append the distance between them to a *sense duplication list* $D_{dup}$. For other nearest neighbors, we append their distances to the *word neighbor list* $D_{nn}$. After populating the two lists, we want to choose a threshold that would prune away all of the sense duplicates while differentiating sense duplications with other distinct neighbor words. Thus, we compute

$$\lambda = \frac{1}{2}(\text{mean}(D_{dup}) + \text{mean}(D_{nn})). \tag{19}$$

| Model | MaxSimC | AvgSimC |
|---|---|---|
| GASI-0.4 | 66.4 | 69.6 |
| GASI-0.4-30K | 64.3 | 69.3 |
| GASI-0.4-post-pruning | 64.6 | 68.9 |

Table 5: Spearman's ranking correlation $100 \times \rho$ on SCWS. GASI-0.4-30K means top 30,000 words are initialized with three senses while the others have one sense.

Table 1 compares the sense embeddings after pruning with the original mode on the Stanford Contextual Word Similarities (SCWS) task Huang et al. (2012). Both AvgSimC and MaxSimC with post-pruning embeddings decrease only a few compare to that from GASI-0.4.

### B.2 Number of Senses vs. Word Frequency

It is a common assumption that more frequent words have more senses. Figure 1 shows a histogram of the number of senses left for words ranked by their frequency, and the results agree with the assumption. Generally, the model learns more sense for high frequent words, except for the most frequent ones. The most frequent words are usually considered stopwords, such as "the", "a" and "our', which have only one common meaning. Moreover, we compare our model initialized with three senses (GASI-0.4, $K = 3$) against the one that has five (GASI-0.4, $K = 5$). Initializing with a larger number of senses, the model is able to uncover more senses for most words.

### B.3 Initialzing $K$ based on Word Frequency

Despite our model has a fixed number of senses. It is easy to implement our model with different number of senses with a mask matrix. And we can define different number of senses for each word based on their frequency. In Table 1, we show the results from a model that only top 30,000 word are initialized with three senses while others have one. The same choice is applied by Neelakantan et al. (2014).