On the Curious Case of ℓ_2 norm of Sense Embeddings

Yi Zhou
University of Liverpool
y.zhou71@liverpool.ac.uk

Danushka Bollegala
University of Liverpool, Amazon
danushka@liverpool.ac.uk

Abstract

We show that the ℓ_2 norm of a static sense embedding encodes information related to the frequency of that sense in the training corpus used to learn the sense embeddings. This finding can be seen as an extension of a previously known relationship for word embeddings to sense embeddings. Our experimental results show that, in spite of its simplicity, the ℓ_2 norm of sense embeddings is a surprisingly effective feature for several word sense related tasks such as (a) most frequent sense prediction, (b) Word-in-Context (WiC), and (c) Word Sense Disambiguation (WSD). In particular, by simply including the ℓ_2 norm of a sense embedding as a feature in a classifier, we show that we can improve WiC and WSD methods that use static sense embeddings.

1 Introduction

Background: Given a text corpus, static word embedding learning methods (Pennington et al. 2014, Mikolov et al. 2013a, etc.) learn a *single* vector (aka *embedding*) to represent the meaning of a word in the corpus. In contrast, static sense embedding learning methods (Loureiro and Jorge 2019a, Scarlini et al. 2020b, etc.) learn *multiple* embeddings for each word, corresponding to the different senses of that word.

Arora et al. (2016) proposed a random walk model on the word co-occurrence graph and showed that if word embeddings are uniformly distributed over the unit sphere, the log-frequency of a word in a corpus is proportional to the squared ℓ_2 norm of the static word embedding, learnt from the corpus. Hashimoto et al. (2016) showed that under a simple metric random walk over words where the probability of transitioning from one word to another depends only on the squared Euclidean distance between their embeddings, the log-frequency of word co-occurrences between two words converges to the negative squared Euclidean distance

measured between the corresponding word embeddings. Mu and Viswanath (2018) later showed that word embeddings are distributed in a narrow cone, hence not satisfying the uniformity assumption used by Arora et al. (2016), however their result still holds for such anisotropic embeddings. On the other hand, Arora et al. (2018) showed that a word embedding can be represented as the linearly-weighted combination of sense embeddings. However, to the best of our knowledge, it remains unknown thus far as to **What is the relationship between the sense embeddings and the frequency of a sense?**, the central question that we study in this paper.

Contributions: First, by extending the prior results for word embeddings into sense embeddings, we show that the **squared ℓ_2 norm of a static sense embedding is proportional to the log-frequency of the sense in the training corpus.** This finding has important practical implications. For example, it is known that assigning every occurrence of an ambiguous word in a corpus to the most frequent sense of that word (popularly known as the Most Frequent Sense (MFS) baseline) is a surprisingly strong baseline for WSD (McCarthy et al., 2004, 2007). Therefore, the theoretical relationship which we prove implies that we should be able to use ℓ_2 norm to predict the MFS of a word.

Second, we conduct a series of experiments to empirically validate the above-mentioned relationship. We find that the relationship holds for different types of static sense embeddings learnt using methods such as GloVe (Pennington et al., 2014) and skip-gram with negative sampling (SGNS; Mikolov et al., 2013b) on SemCor (Miller et al., 1993).

Third, motivated by our finding that ℓ_2 norm of pretrained static sense embeddings encode sense-frequency related information, we use ℓ_2 norm of sense embeddings as a feature for several sense-

related tasks such as (a) to predict the MFS of an ambiguous word, (b) determining whether the same sense of a word has been used in two different contexts (WiC; Pilehvar and Camacho-Collados, 2019), and (c) disambiguating the sense of a word in a sentence (WSD). We find that, regardless of its simplicity, ℓ_2 norm is a surprisingly effective feature, consistently improving the performance in all those benchmarks/tasks. The evaluation scripts is available at: https://github.com/LivNLP/L2norm-of-sense-embeddings.

2 ℓ_2 norm vs. Frequency

Let us first revisit the generative model proposed by Arora et al. (2016) for static word embeddings, where the t-th word, v, in a corpus is generated at step t of a random walk of a *context* vector $\boldsymbol{c}_t$, which represents what is being talked about. The probability, $p(v|c_t)$, of emitting v at time t is modelled using a log-linear word production model, proportionally to $\exp(\boldsymbol{c}_t^\top \boldsymbol{v})$. If $\mathcal{G}_v$ is a *word co-occurrence* graph, where vertices correspond to the words in the vocabulary, $\mathcal{V}$, the random walker can be seen as visiting the vertices in $\mathcal{G}_v$ according to this probability distribution. Arora et al. (2016) showed that the partition function, Z_c, given by (1) for this probabilistic model is a constant Z, independent of the context c.

$$Z_c = \sum_v \exp(\boldsymbol{c}^\top \boldsymbol{v}) \qquad (1)$$

Assuming that the stationary distribution of this random walk is uniform over the unit sphere, Arora et al. (2016) proved the relationship in (2), for d dimensional word embeddings, $\boldsymbol{v} \in \mathbb{R}^d$.

$$\log p(v) = \frac{||\boldsymbol{v}||_2^2}{2d} - \log Z \qquad (2)$$

Let the frequency of v in the corpus be $f(v)$, and the total number of word occurrences be $N = \sum_v f(v)$. $p(v)$ can be estimated using corpus counts as $f(v)/N$. Because N, d, and Z are constants, independent of v, (2) implies a linear relationship between $\log f(v)$ and $||\boldsymbol{v}||_2^2$.

To extend this result to sense embeddings, we observe that the word v generated at step t by the above-described random walk can be uniquely associated with a sense id s_v, corresponding to the meaning of v as used in c_t. If we consider a second *sense co-occurrence* graph $\mathcal{G}_s$, where vertices correspond to the sense ids, then the above-mentioned
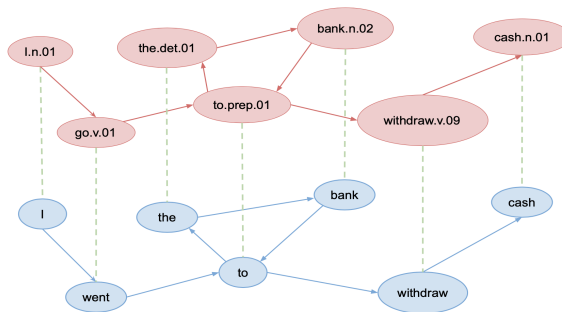

Figure 1: Part of the word co-occurrence graph $\mathcal{G}_v$ (bottom) shown with the corresponding sense co-occurrence graph $\mathcal{G}_s$ (top). Each word in $\mathcal{G}_v$ is mapped to its correct sense in $\mathcal{G}_s$.

corpus generation process corresponds to a second random walk on $\mathcal{G}_s$, as shown in Figure 1.

Although an ambiguous word can be mapped to multiple sense ids across the corpus in different contexts, at any given time step t, a word v is mapped to only one vertex in $\mathcal{G}_s$, determined by the context c_t. Indeed a WSD can be seen as the process of finding such a mapping. The two random walks over word and sense id graphs are isomorphic and converge to the same set of final states (Bauerschmidt et al., 2021). Therefore, an analogous relationship given by (3) can be obtained by replacing word embeddings, $\boldsymbol{v}$, with sense embeddings, $\boldsymbol{s}$, in (2).

$$\log p(s) = \frac{||\boldsymbol{s}||_2^2}{2d_s} - \log Z' \qquad (3)$$

Here, d_s is the dimensionality of the sense embeddings $\boldsymbol{s} \in \mathbb{R}^{d_s}$. Later in § 3, we empirically show that the normalisation coefficient, $Z' = \sum_s \exp(\boldsymbol{c}^\top \boldsymbol{s})$, for sense embeddings also satisfies the self-normalising (Andreas and Klein, 2015) property, thus independent of c. If we abuse the notation $f(s)$ to denote also the frequency of s in the corpus (i.e. the total number of times the random walker visits the vertex s), from (3) it follows that $\log f(s)$ is linearly related to $||\boldsymbol{s}||_2^2$.

3 Empirical Validation

The theoretical analysis described in § 2 implies a linear relationship between $\log f(s)$ and $||\boldsymbol{s}||_2^2$ for the learnt sense embeddings. To empirically verify this relationship, we learn static sense embeddings using GloVe and SGNS from SemCor, which is the largest corpus manually annotated with WordNet (Miller, 1995) sense ids. Specifically, we consider the co-occurrences of senses instead of
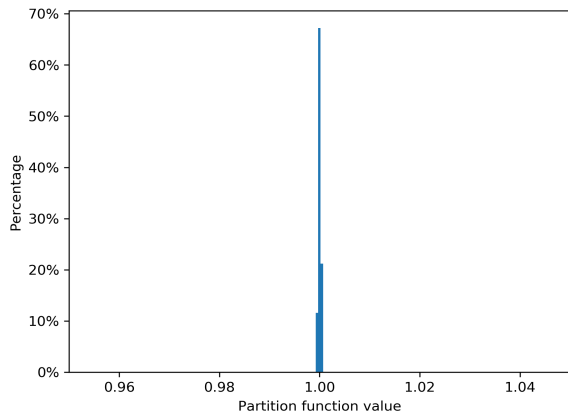
Figure 2: Histogram of the partition function for 1,000 random vectors c for GloVe-sense. The x-axis is normalised by the mean of the values.
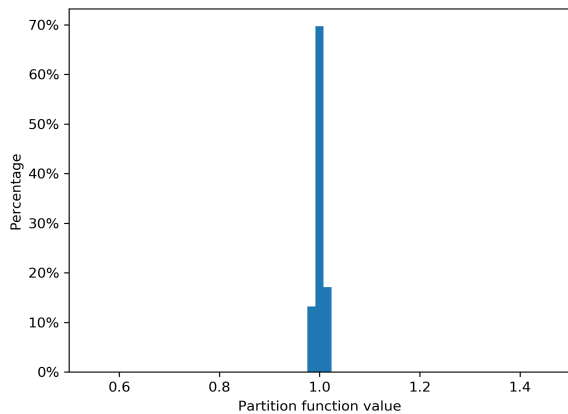

Figure 3: Histogram of partition function for 1,000 random vectors c for SGNS-sense. The x-axis is normalised by the mean of the values.

Figure 4: A linear relationship between $\log f(s)$ (x-axis) and $||s||_2^2$ (y-axis) can be seen for GloVe-sense embeddings represented by the blue dots.

Figure 5: A linear relationship between $\log f(s)$ (x-axis) and $||s||_2^2$ (y-axis) can be seen for SGNS-sense embeddings represented by the blue dots.

words for this purpose. To distinguish the sense embeddings learnt from GloVe and SGNS from their word embeddings, we denote these by respectively **GloVe-sense** and **SGNS-sense**.

Figure 2 shows the partition function for GloVe-sense embeddings. We see that the partition function is tightly concentrated around its mean, showing that sense-embeddings also demonstrate self-normalisation similar to word embeddings. Similar to the histogram for GloVe-sense embeddings, we see that the partition function for SGNS-sense embeddings is also tightly centred around the mean (i.e., 1.0) from Figure 3.

Figure 4 shows the correlation between $\log f(s)$ and $||s||_2^2$ for GloVe-sense. We see a moderate positive correlation (Pearson's $\rho = 0.437$) between these two variables, confirming the linear relationship predicted in §2. Similar to the correlation plot for GloVe-s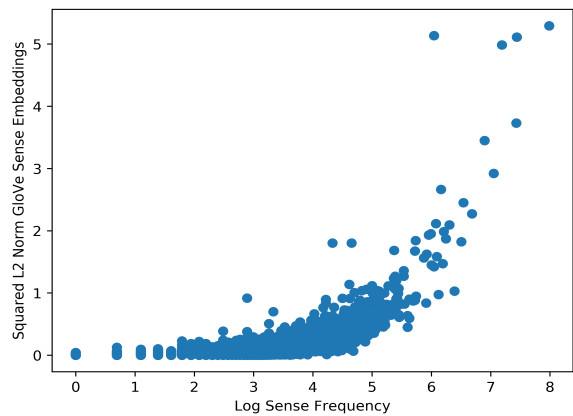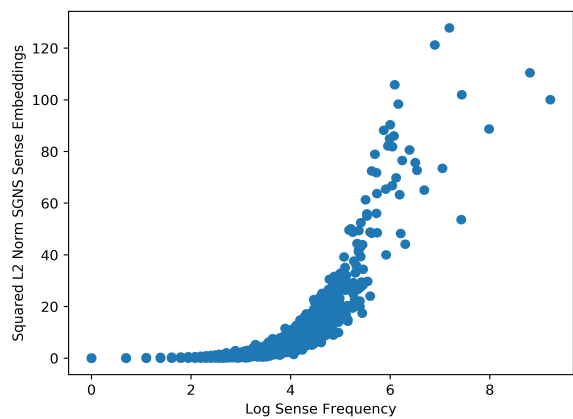ense embeddings, from Figure 5, one can see a positive correlation (Pearson's $\rho = 0.440$) between the log-frequency and squared ℓ_2 norm for the SGNS-sense embeddings.

It is noteworthy however this linear relationship between log-frequency and squared ℓ_2 norm does not hold for contextualised word embeddings such as BERT (Devlin et al., 2019) or static sense embeddings such as LMMS (Loureiro and Jorge, 2019a) that are computed by averaging BERT embeddings (see Appendix B for details). The random walk model described in §2 cannot be applied to contextualised embeddings because the probability of occurrence of a word under the discriminative masked language modelling objectives used to train contextualised word embeddings such as BERT depend on all the words generated before as well as after the target word.

Models	All words	Noun Sample
Random	67.6	26.0
UMFS-WE	73.9	48.0
EnDi	71.4	47.4
WCT-VEC	75.2	48.8
COMP2SENSE	77.9	58.5
Ours		
GloVe-sense with ℓ_2 norm	90.1	92.2
SGNS-sense with ℓ_2 norm	**95.6**	**96.6**

Table 1: Percentage accuracy for the MFS prediction task on SemCor for All Words and the Noun Sample, limited to polysemous nouns. Overall best scores are in bold.

Bins	Max Freq	Min Freq	Word Count
1	15,783	64	545
2	64	32	545
3	32	20	545
4	20	13	545
5	13	9	545
6	9	7	545
7	7	5	545
8	5	3	545
9	3	2	545
10	2	1	545

Table 2: Statistics of each bin of ambiguous words grouped based on their frequency in SemCor.

3.1 Predicting the Most Frequent Sense

To investigate whether frequency of a sense is indeed represented by the squared ℓ_2 norm of its static sense embedding, we conduct an MFS prediction task on SemCor following the setup proposed by Hauer et al. (2019). In this MFS prediction task, given the set of senses of an ambiguous word, we must predict the sense with the highest frequency for that word in SemCor. For this purpose, we filter senses by the lemma and part-of-speech (PoS) of the target word and select the sense with the largest squared ℓ_2 norm using GloVe-sense and SGNS-sense embeddings separately.

In Table 1, we compare our results against a random sense selection baseline and several prior proposals on the MFS benchmark dataset (Hauer et al., 2019). EnDi (Pasini and Navigli, 2018) is a language-independent and fully automatic method for sense distribution learning from raw text. UMFS-WE (Bhingardive et al., 2015) and WCT-VEC (Hauer et al., 2019) both use the distance between word and sense embeddings. COMP2SENSE (Hauer et al., 2019) is a knowledge-based method using WordNet and uses a set of words known as the companions of a target word to determine MFS, based on a sense-similarity function. As seen from Table 1, both GloVe-sense and SGNS-sense outperform all the other methods for **all** words and **noun** sample settings. In particular, for **noun** sample, which contains polysemous nouns that occur at least 3 times in SemCor, both methods obtain more than 35% accuracy improvements over the next best method, providing strong empirical evidence supporting the linear relationship predicted by (3).

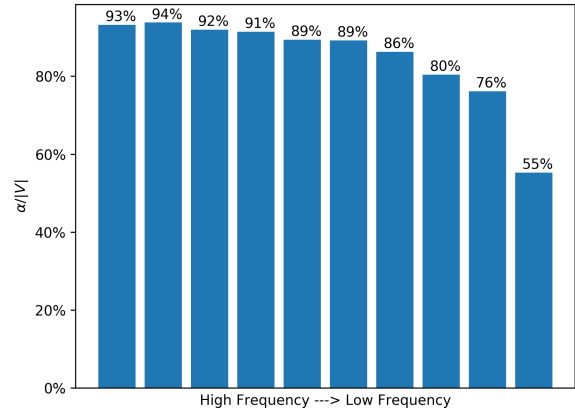If the ℓ_2 norm of a sense embedding relates to the frequency of that sense, the ℓ_2 norm of the most

Figure 6: The trend of $\alpha/|\mathcal{V}|$ from high frequent words to low frequent words.

frequent sense should be always greater than the ℓ_2 norm of the next frequent sense of an ambiguous word. To further investigate this, we sort the set of ambiguous words in SemCor based on their frequency and divide them into 10 subsets (i.e., bins). The summary statistics of each subset is shown in Table 2. For each ambiguous word w, we find its most frequent sense w_m and next frequent sense w_n in SemCor. Note that w_m and w_n are determined based on their frequency in SemCor and not according to how they are sorted in the WordNet.

Let us denote the ℓ_2 norms of w_m and w_n by $||\boldsymbol{w_m}||_2$ and $||\boldsymbol{w_n}||_2$ respectively, and $\alpha = \sum_{w \in \mathcal{V}} \mathbb{I}(||\boldsymbol{w_m}||_2 > ||\boldsymbol{w_n}||_2)$, where $\mathbb{I}(x)$ is the indicator function that returns 1 if x is True and 0 otherwise. $\mathcal{V}$ is the set of ambiguous words (i.e., the words that have a least two distinct senses in SemCor). We compute the percentage $\alpha/|\mathcal{V}|$ for the total vocabulary and shown the result in Figure 6. We observe that the second bin obtains the highest $\alpha/|\mathcal{V}|$ score over all the bins. Moreover, the $\alpha/|\mathcal{V}|$ scores decrease with the frequency of the ambiguous words. This result indicates that

Models	Test
LMMS-based	
LMMS (Loureiro and Jorge, 2019a)	64.8
LMMS + ℓ_2 norm of GloVe-sense	65.8
LMMS+ℓ_2 norm of SGNS-sense	**67.0**
ARES-based	
ARES (Scarlini et al., 2020b)	66.6
ARES+ℓ_2 GloVe-sense	66.6
ARES+ℓ_2 SGNS-sense	66.7

Table 3: Accuracies on the WiC test sets for LMMS- (top) and ARES- (bottom) based classifiers. Overall best score is in bold.

the relationship between the ℓ_2 norm of a sense embedding and its frequency is stronger for high frequent words than low frequent ones.

3.2 Predicting Word Sense in Context

We evaluate the ℓ_2 norm of sense embeddings in WiC and WSD as downstream tasks. In WiC, given an ambiguous word w occurring in two contexts c_1 and c_2, we must predict whether w occurs in c_1 and c_2 with the same sense. We follow Loureiro and Jorge (2019b), and train a binary logistic regression model on WiC training set using different sets of similarities between static sense embeddings and contextualised embeddings obtained from a language model (i.e. BERT) as features. Further details regarding the features used for this classifier are given in §A.3. We consider two current state-of-the-art sense embeddings, LMMS and ARES (Scarlini et al., 2020b), and include ℓ_2 norm of static sense embeddings as extra features, and measure the gain in performance.

From Table 3 we see that by including ℓ_2 norm of GloVe-sense and SGNS-sense embeddings as features, we obtain more than 1% gains in accuracy over the original LMMS on WiC test sets. ARES+ℓ_2 norm GloVe-sense obtains the same score as the ARES baseline, while ARES+ℓ_2 norm SGNS-sense achieves a slight improvement on the test set. This shows that ℓ_2 norm of static sense embeddings encodes sense frequency related information, which improves the performance in WiC when used with static sense embeddings. This is noteworthy given that ℓ_2 norm is a single feature compared to LMMS and ARES, which are both 2048 dimensional.

Methods	SE2	SE3	SE07	SE13	S15	ALL
LMMS-based						
LMMS	76.3	75.6	68.1	75.1	77.0	75.4
LMMS+ℓ_2 norm GloVe-sense	77.8	76.9	70.5	76.6	77.8	76.8
LMMS+ℓ_2 norm SGNS-sense	77.5	77.4	69.7	77.1	78.1	76.9
ARES-based						
ARES	78.0	77.1	71.0	77.3	**83.2**	77.9
ARES+ℓ_2 norm GloVe-sense	**78.4**	**77.8**	**71.6**	77.9	82.4	**78.3**
ARES+ℓ_2 norm SGNS-sense	77.6	77.5	68.6	**78.0**	82.0	77.7

Table 4: F1 on the test sets of the all-words English WSD framework for LMMS- (top) and ARES- (bottom) based method. Overall best scores are in bold.

3.3 Word Sense Disambiguation

We further evaluate ℓ_2 norm of static sense embeddings using the English all-words WSD framework (Raganato et al., 2017). For this purpose, we train a binary logistic regression classifier using the two features – (a) the similarity between the contextualised embedding and a sense embedding of the target word, and (b) the squared ℓ_2 norm of the sense embedding. We use SemCor training data and consider the correct sense of the target word as a positive instance, and its other senses as negative instances. At inference time, we predict the sense with the highest probability of being positive as the correct sense of the test word in the given context. Likewise in the WiC evaluation in §3.2, we measure the improvements in performance over LMMS and ARES, with using ℓ_2 norm as a feature for WSD.

Table 4 shows the F1 scores for all-words English WSD datasets. ARES+ℓ_2 norm Glove-sense reports the best performance in three out of the five datasets, and obtains the best performance on ALL (i.e., concatenation of all the test sets), whereas ARES+ℓ_2 norm SGNS-sense reports the best performance in SE13. In LMMS-based evaluations, we see that always either one or both GloVE/SGNS-sense ℓ_2 norms improve over the vanilla LMMS. This shows that we are able to improve the performance of both LMMS and ARES by simply adding ℓ_2 norm of static sense embeddings as extra features.

4 Conclusion

We investigated log-frequency and ℓ_2 norm of sense embeddings, and found a linear relationship between those. Experimental results indicate that, despite its simplicity, ℓ_2 norm of sense embedding is a surprisingly effective feature for MFS prediction, WiC and WSD tasks.

5 Limitations

This paper makes both theoretical and empirical contributions related to sense embeddings. In this section, we highlight some of the important limitations in terms of both theory and empirical evaluations we made in the paper. We hope this will be useful when extending our work in the future by addressing these limitations.

On the theoretical side, as we already stated in §3, the generative random walk model is not applicable to contextualised word embeddings obtained by a language model such as BERT. As shown in Appendix B, although the partition function for BERT demonstrates the self-normalising property when word embeddings are computed by averaging the context embeddings of a word across a corpus, the squared ℓ_2 norm of these BERT-based word does *not* demonstrate a linear relationship with the logarithm of the frequency of that word in the corpus. This has important consequences with regard to static sense embedding methods such as LMMS and ARES, which also use BERT to obtain sense representations from dictionaries such as the WordNet or sense-labelled corpora such as SemCor. Although LMMS and ARES are full-coveraged state-of-the-art static sense embeddings, their ℓ_2 norms does not satisfy the linear relationship, which we derived in this paper due to this reason as shown in Appendix B. An important future research direction would be to develop a random walk model for masked language models such as BERT, and analytically derive a relationship between the word embeddings and frequency of words. We also exclude contextualised sense embedding methods such as SenseBERT (Levine et al., 2020) from our analysis due to the same reason.

On the empirical side, a limitation of our evaluation is that it is limited to the English language. There are WSD and WiC benchmarks for other languages such as SemEval-13, SemEval-15, XL-WSD (Pasini et al., 2021) and WiC-XL (Raganato et al., 2020), as well as multilingual sense embeddings such as ARES$_m$ (Scarlini et al., 2020b) and SensEmBERT (Scarlini et al., 2020a). Extending our evaluations to cover multilingual sense embeddings is deferred to future work.

6 Ethical Considerations

In this paper, we inspect the relationship between ℓ_2 norm of static sense embedding and its frequency in the training corpus. We evaluate the effectiveness of ℓ_2 norm of static sense embeddings on several experiments, i.e., MFS prediction, WiC and WSD tasks. In particular, we did not annotate any datasets by ourselves in this work and used multiple corpora and benchmark datasets that have been collected, annotated and repeatedly used for evaluations in prior works. To the best of our knowledge, no ethical issues have been reported concerning these datasets. However, we note that it has been reported that pretrained sense embeddings encode various types of social biases such as gender and racial biases (Zhou et al., 2022). Our experiments show that including ℓ_2 norm of sense embeddings to be an effective strategy for improving the performance in sense-related tasks such as WSD and WiC. However, it remains an open question whether ℓ_2 norm also encodes social biases, and if so how to mitigate those biases from affecting downstream tasks.

References

Jacob Andreas and Dan Klein. 2015. When and why are log-linear models self-normalizing? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 244–249, Denver, Colorado. Association for Computational Linguistics.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.

Roland Bauerschmidt, Tyler Helmuth, and Andrew Swan. 2021. The geometry of random walk isomorphism theorems. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 57(1).

Sudha Bhingardive, Dhirendra Singh, V Rudramurthy, Hanumant Harichandra Redkar, and Pushpak Bhattacharyya. 2015. Unsupervised most frequent sense detection using word embeddings. In *HLT-NAACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technolo-*

gies, Volume 1 (Long and Short Papers), pages 4171–4186.

Tatsunori Hashimoto, David Alvarez-Melis, and Tommi Jaakkola. 2016. Word embeddings as metric recovery in semantic spaces. *Transactions of the Association for Computational Linguistics*, 4:273–286.

Bradley Hauer, Yixing Luan, and Grzegorz Kondrak. 2019. You shall know the most frequent sense by the company it keeps. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 208–215. IEEE.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online. Association for Computational Linguistics.

Daniel Loureiro and Alipio Jorge. 2019a. Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy.

Daniel Loureiro and Alipio Jorge. 2019b. Liaad at semdeep-5 challenge: Word-in-context (wic). In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 1–5.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553 – 590.

Diana McCarthy, Rob Koeling, Julie Weeds, and John A Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 279–286.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv: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 NIPS*, pages 3111–3119.

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

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Tommaso Pasini and Roberto Navigli. 2018. Two knowledge-based methods for high-performance sense distribution learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13648–13656.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8758–8765.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020b. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539, Online.

Yi Zhou, Masahiro Kaneko, and Danushka Bollegala. 2022. Sense embeddings are also biased – evaluating social biases in static and contextualised sense embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1924–1935, Dublin, Ireland. Association for Computational Linguistics.

Appendix

A Experimental Setup

In this section, we report the setup of experiments we conduct to evaluate the effectiveness of ℓ_2 norm of static sense embeddings.

A.1 Training GloVe-sense and SGNS-sense

We train our GloVe-sense and SGNS-sense on SemCor training data. Specifically, for each target word w in a context c, we train a vector and assign the annotated sense label to it. For GloVe-sense, we use its Python-based implementation.[1] We set the co-occurrence window to 10 tokens, number of dimensions to 300 and the initial learning rate to 0.05 for the vanilla stochastic gradient descent. We train the embeddings for 30 epochs with 2 parallel threads. To train SGNS-sense, we use the Word2Vec module from gensim.models.[2] We set the min_count to 1 and the dimensionality of the embeddings to 300, and the remainder of the hyperparameters remain at their default values.

A.2 Predicting the Most Frequent Sense

We conduct our MFS experiment on SemCor. Given a target word w in a context c, we first select a set of candidate senses based on w's lemma and PoS. Then we compute the ℓ_2 norm of the static sense embedding for each sense in the candidate set. Finally, we take the sense with the maximum ℓ_2 norm score as the predicted MFS for w. Then we compare our prediction with the MFS of w according to the sense occurrence in SemCor, and compute the accuracy scores.

A.3 Word Sense Disambiguation

We consider the Word Sense Disambiguation task as a binary classification problem and train a Logistic Regression binary classifier on SemCor. To evaluate the baselines, i.e., LMMS (LMMS SP-WSD: sensekeys[3]) and ARES on WSD, given a

word w in a sentence c, we first compute its contextualised embedding using BERT (bert-large-cased) model by averaging the last four layers, denoted by $t(w, c)$. We then compute the cosine similarity between $t(w, c)$ and the sense embedding $s(w)$ corresponding to the senses of w based on WordNet as a feature. We use the binary logistic regression classifier implemented in sklearn with the default parameters. For our proposed method, we simply append the ℓ_2 norm of static sense embedding of w as an additional feature. To avoid any discrepancies in the scoring methodology, we use the English all-words WSD framework (Raganato et al., 2017) and its offical scoring scripts.

A.4 Predicting Word Sense in Context

We train a binary logistic regression classifier[4] on the WiC training set. Following the work from Loureiro and Jorge (2019b), we compute four similarities between sense and contextualised embeddings, and consider those as features. Specifically, given a target word w in two contexts c_1 and c_2, similar to § 3.3, we first determine the sense-specific embeddings for w in c_1 and c_2, denoted by $s_1(w)$ and $s_2(w)$. Then we use the cosine similarities between the two vectors in the following four pairs as features, requiring no expensive fine-tuning procedure: $(s_1(w),\ s_2(w))$, $(t(w, c_1),\ t(w, c_2))$, $(s_1(w),\ t(w, c_1))$, $(s_2(w),\ ,\ t(w, c_2))$. Contextualised embeddings are not ℓ_2 normalised in this experiment. Here again, similar to the WSD settings described above, with respect to our proposed method, we simply append the ℓ_2 norm of the static sense embedding of w as the fifth feature.

B Static Sense Embeddings from Contextualised Word Embeddings

We investigate whether the self-normalising and linearity properties hold for contextualised embeddings obtained from language models. For this purpose, we compute the static word embeddings for the words appearing in SemCor using contextualised embeddings learnt by BERT. Specifically, we compute the average over the contextualised BERT embeddings for all of the occurrences of a word in SemCor, and consider it as the static (i.e. context-independent) BERT embedding for that word. To distinguish the contextualised embeddings learnt

---

[1] https://github.com/maciejkula/glove-python
[2] https://radimrehurek.com/gensim/models/word2vec.html
[3] https://github.com/danlou/LMMS

[4] We use the default parameters in scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.
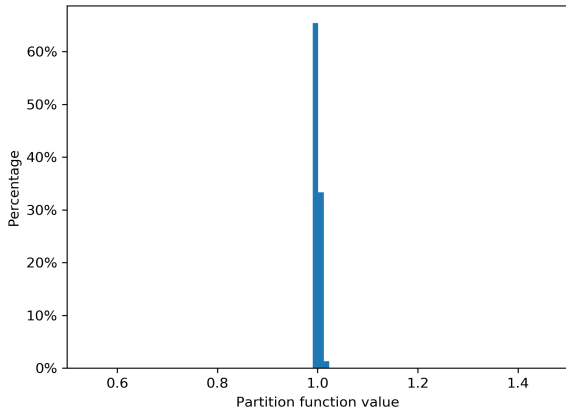
Figure 7: Histogram of the partition function for 1,000 random vectors c for BERT-word. The x-axis is normalised by the mean of the values.
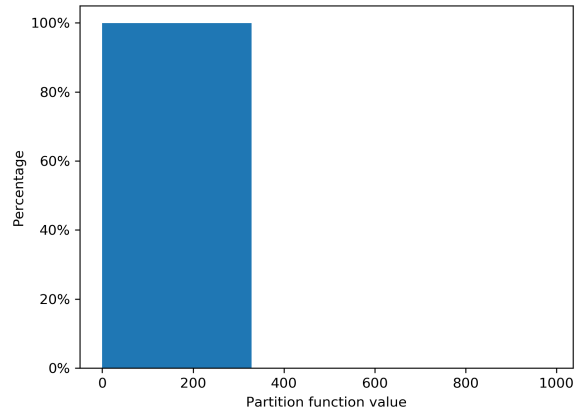

Figure 9: Histogram of partition function for 1,000 random vectors c for LMMS$_{sc}$. The x-axis is normalised by the mean of the values.
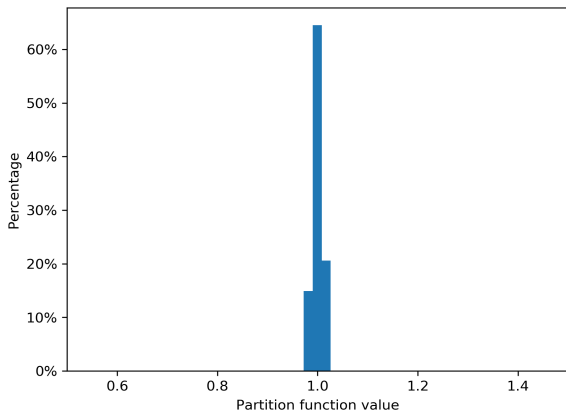

Figure 8: Histogram of the partition function for 1,000 random vectors c for LMMS. The x-axis is normalised by the mean of the values.
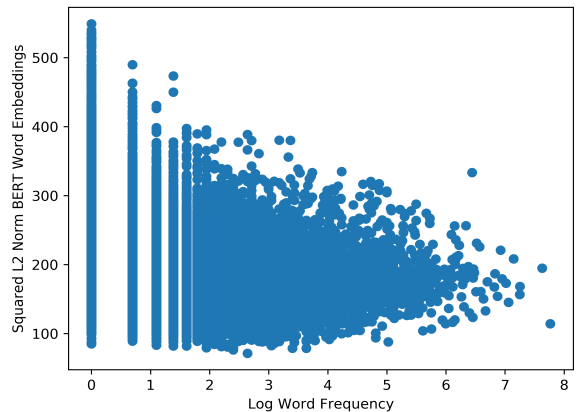

Figure 10: The linear relationship between the squared ℓ_2 norms of BERT-word embeddings and the logarithms of the sense frequencies. Each dot in the plot corresponds to a word, where x-axis is the natural logarithm of the sense frequency, and y-axis is the squared ℓ_2 norm of the word embedding. The Pearson correlation coefficient between the two is -0.316.

from BERT, we name the static BERT embeddings as BERT-static in the remainder of this paper.

Recall that LMMS uses BERT to compute sense embeddings from SemCor and WordNet's glosses. Therefore, if BERT-static satisfies the self-normalising and linearity properties described in § 2, LMMS embedding must satisfy these properties as well. In addition, we take the first step of LMMS training procedure from the work of Loureiro and Jorge (2019a)[5] and train static sense embeddings only on SemCor data without normalising the learnt sense embeddings (doing so would remove ℓ_2 norm related information from the sense embeddings). To differentiate this version of LMMS embeddings from the full-coverage LMMS embeddings, we refer to it as LMMS$_{sc}$ (here, sc stands for SemCor). We then test if the self-normalising and linearity properties hold for

BERT-static, LMMS and LMMS$_{sc}$.

Figure 7, 8 and 9 show the histogram of partition functions for BERT-word, LMMS and LMMS$_{sc}$, respectively. We observe that the histograms of both BERT-static and LMMS are centred around mean, while LMMS$_{sc}$ is not. This shows that LMMS$_{sc}$ does not satisfy self normalising, while BERT-static and LMMS do.
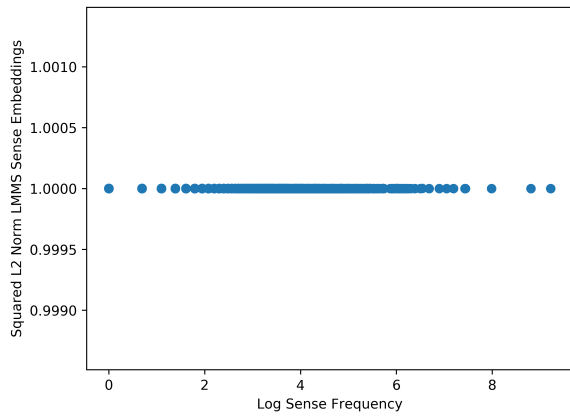
2601

Figure 11: The linear relationship between the squared ℓ_2 norms of LMMS embeddings and the logarithms of the sense frequencies. Each dot in the plot corresponds to a word, where x-axis is the natural logarithm of the sense frequency, and y-axis is the squared ℓ_2 norm of the word embedding. The Pearson correlation coefficient between the two is -0.005.
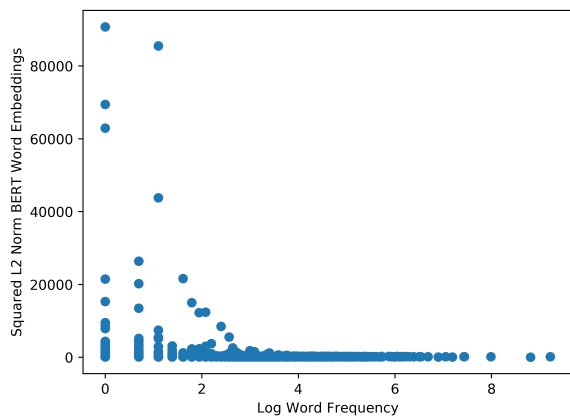

Figure 12: The linear relationship between the squared ℓ_2 norms of LMMS$_{sc}$ embeddings and the logarithms of the sense frequencies. Each dot in the plot corresponds to a word, where x-axis is the natural logarithm of the sense frequency, and y-axis is the squared ℓ_2 norm of the word embedding. The Pearson correlation coefficient between the two is -0.010.

Figure 10, 11 and 12 show the correlation between squared ℓ_2 norms of the word/sense embeddings and the logarithms of sense/word frequencies for BERT-static, LMMS and LMMS$_{sc}$, respectively. From the figures we see that none shows a linear relationship. This indicates that sense frequency related information is not encoded in the ℓ_2 norm of LMMS (or BERT) embeddings.