

Dirichlet-Smoothed Word Embeddings for Low-Resource Settings

Jakob Jungmaier, Nora Kassner, Benjamin Roth

Center for Information and Language Processing, Ludwig-Maximilians-Universität München
jakob.jungmaier@campus.lmu.de, {kassner, beroth}@cis.uni-muenchen.de

Abstract

Nowadays, classical count-based word embeddings using positive pointwise mutual information (PPMI) weighted co-occurrence matrices have been widely superseded by machine-learning-based methods like word2vec and GloVe. But these methods are usually applied using very large amounts of text data. In many cases, however, there is not much text data available, for example for specific domains or low-resource languages. This paper revisits PPMI by adding Dirichlet smoothing to correct its bias towards rare words. We evaluate on standard word similarity data sets and compare to word2vec and the recent state of the art for low-resource settings: Positive and Unlabeled (PU) Learning for word embeddings. The proposed method outperforms PU-Learning for low-resource settings and obtains competitive results for Maltese and Luxembourgish.

Keywords: Statistical and Machine Learning Methods, Language Modelling, Neural Language Representation Models, Semantics, Less-Resourced/Endangered Languages

1. Introduction

Word embeddings that capture the meaning of words as vectors are the basis of much of the recent progress in natural language processing. Nowadays, the classical count-based way to obtain such word embeddings from a corpus by using positive pointwise mutual information (PPMI) weighted co-occurrence matrices has been widely superseded by machine-learning-based methods like word2vec (Mikolov et al., 2013a; Mikolov et al., 2013b) and GloVe (Pennington et al., 2014). These methods are usually applied using very large amounts of text data. But in many settings there is not much text data available, for example for specific domains or low-resource languages.

Recent investigations by Konovalov and Tumunbayarova (2018) and Jiang et al. (2018) give rise to the conjecture that variants of the classical method to compute word embeddings might be more suitable for smaller corpora than the machine-learning-based methods. Additionally, Levy et al. (2015) suggest that certain choices of system design and hyperparameters for computing word embeddings can even be more impactful than the choice of the underlying model architecture itself.

In the spirit of these results, we propose to use a variant of the classical method for low-resource languages which tries to address a well-known problem of the PMI measure: its bias towards rare words (Turney and Pantel, 2010; Levy et al., 2015; Jurafsky and Martin, 2018). Following an idea mentioned by Turney and Pantel (2010) and Jurafsky and Martin (2018), we apply *Dirichlet smoothing* to weaken PMI’s bias towards rare words and compare it to the word2vec skip-gram with negative sampling (SGNS) and Positive and Unlabeled (PU) Learning for word embeddings by Jiang et al. (2018).

Models are trained on the English enwik9 corpus and evaluated using standard word similarity data sets. We investigate the models’ performance for varying corpus size. In a case study, to demonstrate their practicability for low-resource languages, the three methods are used to train word embeddings for Maltese and Luxembourgish using respective Wikipedia dumps.

We show that our method i) outperforms strong baseline methods trained on enwik9 with biggest advantage on smaller corpora ii) scores best on the RW (Rare Word) similarity data set (Luong et al., 2013) and iii) obtains outperforming and competitive results in the low-resource setting for Maltese and Luxembourgish.

We make our code publicly available.¹

2. Method

Our proposed method adds Dirichlet smoothing to PPMI embeddings. To compute the standard PPMI embeddings, first, the corpus is scanned in windows consisting of a middle word and its surrounding context words and a word co-occurrence matrix is computed. Words that are frequent in the corpus cause high co-occurrence counts while more infrequent words (that might even be more informative) only result in very low counts. To prevent the overestimation of frequent words, the raw co-occurrence counts are substituted by *pointwise mutual information (PMI)* (Church and Hanks, 1990). The PMI of a middle word w and a context word c is calculated as follows:

$$\text{PMI}(w, c) = \log \left(\frac{P(w, c)}{P(w)P(c)} \right) \quad (1)$$

where $P(w, c)$ is the probability that w and c occur in the same context (window) and $P(w)$ and $P(c)$ are the probabilities of the independent occurrences of the single words w and c . These probabilities are estimated from the co-occurrence matrix by maximum likelihood estimation.

The numerator in equation (1) is the probability of a co-occurrence of the words given their actual distribution in the corpus (and some fixed window size), while the denominator is the expected probability that the two words co-occur given they were distributed independently across the corpus.

PMI faces a problem if two words do not co-occur at all. Then, the fraction will be zero and consequently

¹<https://github.com/jungmaier/dirichlet-smoothed-word-embeddings>

$\text{PMI}(w, c) = \log(0) = -\infty$. A rather pragmatic solution is to simply leave these entries out and let them stay implicit zeros. But this results in an inconsistent matrix since it attributes higher correlation (namely zero) to word pairs that never appeared than it does to word pairs that merely appeared less often than expected (negative values). However, it is still a convenient way of weighting a co-occurrence matrix that has proven to work well in practice (Levy and Goldberg, 2014).

Bullinaria and Levy (2007) demonstrated that the performance of the resulting vectors for word similarity tasks is better if all negatively correlated values are excised from the matrix. The result is called *positive* pointwise mutual information:

$$\text{PPMI}(w, c) = \max(\text{PMI}(w, c), 0). \quad (2)$$

The reason why PPMI performs better than PMI with respect to word similarity might be that usually the fact that two words appear less often together than expected does not convey much information. After all, even for humans it is very hard to name words that are negatively correlated, while it is relatively easy to name words that are positively correlated (Levy and Goldberg, 2014).

(P)PMI suffers from a bias towards rare words (Turney and Pantel, 2010; Levy et al., 2015; Jurafsky and Martin, 2018). The original intention to use PMI is to reduce the influence of the absolute frequency of words in the corpus, which also means to lower the weight of a co-occurrence with a more frequent word in comparison to the co-occurrence with a less frequent word. But one effect of this weighting procedure is that co-occurrences with rare words result in very high PMI values, which in turn overestimates the influence of co-occurrences with rare words.

Turney and Pantel (2010) mention the following case: suppose that two words w and c are strongly statistically dependent, i.e., $P(w, c) \approx P(w) \approx P(c)$. For example, this could be the case for a collocation like “San Francisco”. Then, the PMI value of w and c is $\log(1/P(w))$ and consequently this value increases when the probability of the word w decreases. Now, supposed that “San Francisco” only appears, say, once in the corpus, the PMI value will be very high. But that in turn will skew the relation to the PMI values of other co-occurrences. As Levy et al. (2015) point out, this even creates a situation in which the top “distributional features” of a word, i.e., its context words, are often extremely rare words. But these do not necessarily appear in the respective representations of words that are semantically similar to that word. Moreover, the chance for the co-occurrence with a very rare word to be rather accidental than systematic is certainly higher than for co-occurrences with more frequent words. This means that (P)PMI might give very high weights to words that are in fact not significantly connected.

We approach the problem of rare words for PMI building on an idea mentioned by (Turney and Pantel, 2010) and (Jurafsky and Martin, 2018), which is to use *Dirichlet smoothing*. Usually, Dirichlet smoothing is used to obtain non-zero probabilities for unseen events by adding a small pseudo count λ in each likelihood estimation. For the present purpose, the idea is to add a small pseudo count to every entry

in the co-occurrence matrix. Since the same λ will be added to every entry, the probability of frequent co-occurrences will be lowered a bit while that of rare co-occurrences will be raised. The same will happen to the probabilities of single words.

To maintain the sparsity of the matrix, a smoothed version of PMI is computed for the non-zero entries only. After all, it is not the aim here to get probabilities for co-occurrences that did not appear in the training corpus.

Following this idea, the smoothed PMI_λ will be computed as follows:

$$\text{PMI}_\lambda(w, c) = \log\left(\frac{P_\lambda(w, c)}{P_\lambda(w)P_\lambda(c)}\right) \quad (3)$$

where

$$P_\lambda(w, c) = \frac{f(w, c) + \lambda}{\sum_{(w', c') \in V_w \times V_c} f(w', c') + \lambda|V_w \times V_c|} \quad (4)$$

where $f(w, c)$ denotes the frequency of the co-occurrence of w and c , V_w the vocabulary of all middle words, and V_c the vocabulary of all context words. To preserve a probability distribution, the counts also have to be adjusted for the probability calculations of the single words w and c :

$$P_\lambda(w) = \frac{\sum_{c \in V_c} f(w, c) + \lambda|V_c|}{\sum_{(w', c') \in V_w \times V_c} f(w', c') + \lambda|V_w \times V_c|} \quad (5)$$

$$P_\lambda(c) = \frac{\sum_{w \in V_w} f(w, c) + \lambda|V_w|}{\sum_{(w', c') \in V_w \times V_c} f(w', c') + \lambda|V_w \times V_c|}. \quad (6)$$

Here, the λ 's in the nominators have to be multiplied by $|V_w|$ (respectively $|V_c|$) since adding λ to every co-occurrence count in the matrix will raise the count of the single words by $\lambda|V_w|$ (respectively $\lambda|V_c|$). Using these formulas it is now possible to “pretend” to add λ to every count while in fact only computing PMI_λ for the existing counts. For the actual weighting of the co-occurrence matrix, only the positive PMI_λ -values will be used, which yields PPMI_λ .

The PPMI_λ matrix is still a very large but sparse matrix. To obtain dense word embeddings with only few dimensions, we follow the usual way of using *truncated singular value decomposition* (SVD) for dimensionality reduction:

$$\mathbf{X}' = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T \quad (7)$$

where the original matrix \mathbf{X} is decomposed into orthogonal unit-length column matrices \mathbf{U} , \mathbf{V} , and the diagonal matrix $\mathbf{\Sigma}$ of ordered singular values of \mathbf{X} (Deerwester et al., 1990; Levy and Goldberg, 2014). For the reduced matrix \mathbf{X}' only the largest k singular values in $\mathbf{\Sigma}$ and the corresponding columns of \mathbf{U} and \mathbf{V} are considered. We follow Levy et al. (2015) and use \mathbf{U}_k as our final word embedding matrix and dismiss $\mathbf{\Sigma}_k$.

We refer to the full approach from corpus to word embeddings as SVD-PPMI_λ in this paper.

3. Experimental Setup

We compared the proposed method SVD-PPMI_λ to two baseline methods: word2vec SGNS (Mikolov et al., 2013a; Mikolov et al., 2013b) and PU-Learning for word embeddings (Jiang et al., 2018).

3.1. Corpora

First, we train SVD-PPMI λ , word2vec SGNS, and PU-Learning on *enwik9* which consists of the first 10⁹ bytes of an English Wikipedia dump from 2006, provided for download by Matt Mahoney.² Removal of the markup language and further preprocessing was done by a Perl script by Mahoney to be found on the same web page. Punctuation marks were removed completely, all letters were lowercased and words tokenized by whitespace. The resulting text file contains 124,301,827 tokens in total with a vocabulary of 833,185 words.

To see the influence of the corpus size on the quality of the trained word embeddings, we used differently sized subsets of the *enwik9* corpus. For the following experiments, we used the first 1, 2, 4, 8, 16, 32, and 64 million words.

Experiments for Luxembourgish and Maltese were conducted using dumps of the Luxembourgish and Maltese Wikipedias from 2019. The complete dumps were preprocessed in the same manner as for the *enwik9* corpus. The resulting corpus for Luxembourgish contains 6,268,907 tokens with a vocabulary size of 283,168. The resulting corpus for Maltese contains 1,617,402 tokens with a vocabulary size of 87,902.

3.2. Evaluation

The final word embeddings are evaluated on five word similarity data sets: *RG-65* (Rubenstein and Goodenough, 1965), *WordSim-353* (Finkelstein et al., 2002), *SimLex-999* (Hill et al., 2015), *MEN* (Bruni et al., 2014), and the *RW* (*Rare Word*) data set by Luong et al. (2013). Each of these data sets contains a number of word pairs with a corresponding gold standard similarity score assigned by human annotators. For every word pair in a data set, the cosine similarity of the corresponding word embeddings is computed. For the case that some word occurs in a word pair for which no corresponding word embedding was trained, the similarity for the pair will be set to zero. The final score is Spearman’s rank correlation coefficient (Spearman’s ρ) of the scores assigned by humans and the cosine similarities of the corresponding word embeddings.

In a similar manner as Jiang et al. (2018) obtain data sets for languages other than English, data sets for Luxembourgish and Maltese were obtained for *RG-65*, *WordSim-353*, *SimLex-999*, and *MEN* via the Google translation API.³ Word pairs containing multi-word expressions after translation were removed. Occasionally, pairs with very similar words in the original English data sets resulted in translated pairs of twice the same word. Respective pairs were discarded as well. Manual inspection of the final word pairs indicates that a small proportion of terms seem to remain English after translation, but this is not considered to be a problem since, after all, the conditions are equal for each approach tested here.⁴ Table 1 shows the used data sets and their respective sizes.

²www.cs.fit.edu/~mmahoney/compression/textdata.html

³<https://cloud.google.com/translate>

⁴However, since this effect seemed to be particularly pronounced for the *RW* data set, no translations of this data set were used here.

Data set	English	Luxembourgish	Maltese
<i>RG-65</i>	65	61	62
<i>WordSim-353</i>	353	342	338
<i>SimLex-999</i>	999	937	960
<i>MEN</i>	3000	2904	2852
<i>RW</i>	2034	—	—

Table 1: Overview of the used word similarity data sets and their sizes (number of word pairs).

3.3. Implementation and Hyperparameter Choices

For the comparison with the two baseline methods, the original implementations by the authors were used, i.e., for word2vec SGNS the original C-implementation by Mikolov et al.⁵ and for PU-Learning for word embeddings the original code provided by Jiang et al.⁶

For all compared methods the window size was set to 5, and the minimum count for words was set to 1, i.e., representations were trained for all words in the corpora. The length of all word embeddings was set to the default word2vec dimension of 100.

For the model-specific hyperparameters of the baseline methods the respective default values of the provided implementations were used. The λ -parameter of our own model was selected based on the comparison of different values described in section 4.1.

4. Results and Discussion

Table 2 summarizes Spearman’s rank correlation coefficient for all similarity test sets. Trained on *enwik9*, our proposed method outperforms the two baseline methods on all similarity test sets.

In the low-resource setting, SVD-PPMI λ outperforms baselines for Maltese and shows competitive performance for Luxembourgish embeddings.

4.1. λ -Parameter

The influence of the λ -parameter on the performance of SVD-PPMI λ is shown by a comparison of the average performance on all five word similarity data sets. λ is increased from 10⁻⁶ to 1. To see the influence of the corpus size and to select the best λ ’s for the following experiments, this comparison was made for the *last* 1, 2, 4, 8, 16, 32, and 60 million tokens of *enwik9*. The results can be seen in figure 1. It shows that a good choice for λ can increase the performance by around 10 % (for the last 8M tokens) to 16 % (for the last 2M tokens). A λ chosen too high on the other hand can drop the performance considerably, as can be seen in the case of $\lambda \approx 0.3$ for the last 2M tokens, which decreases the score by around 27 %. However, in most cases it seems that a λ chosen too low does not cause a notable drop in performance. This yields that, if in doubt, it might be advisable to stick to a smaller λ . For increasing corpus size, the optimal λ -value also has the tendency to increase.

⁵<https://github.com/tmikolov/word2vec>

⁶<https://github.com/uclanlp/PU-Learning-for-Word-Embedding>

Language	Method	RG-65	WordSim-353	SimLex-999	MEN	RW
English	SGNS	.718	.674	.306	.690	.294
	PU-Learning	.700	.689	.267	.718	.217
	SVD-PPMI $_{\lambda}$.744	.738	.314	.726	.308
Luxembourgish	SGNS	.146	.133	.069	.186	—
	PU-Learning	.267	.239	.075	.337	—
	SVD-PPMI $_{\lambda}$.302	.193	.054	.351	—
Maltese	SGNS	.063	.047	.074	.168	—
	PU-Learning	.059	.191	.068	.320	—
	SVD-PPMI $_{\lambda}$.141	.208	.106	.351	—

Table 2: Comparison of SGNS (word2vec), PU-Learning, and SVD-PPMI $_{\lambda}$ (ours) concerning word similarity tasks for English, Luxembourgish, and Maltese. Spearman’s ρ between the human annotated gold standard and the cosine similarity scores are shown. Embeddings are trained on enwik9 and 2019 dumps of the Luxembourgish and Maltese Wikipedias respectively.

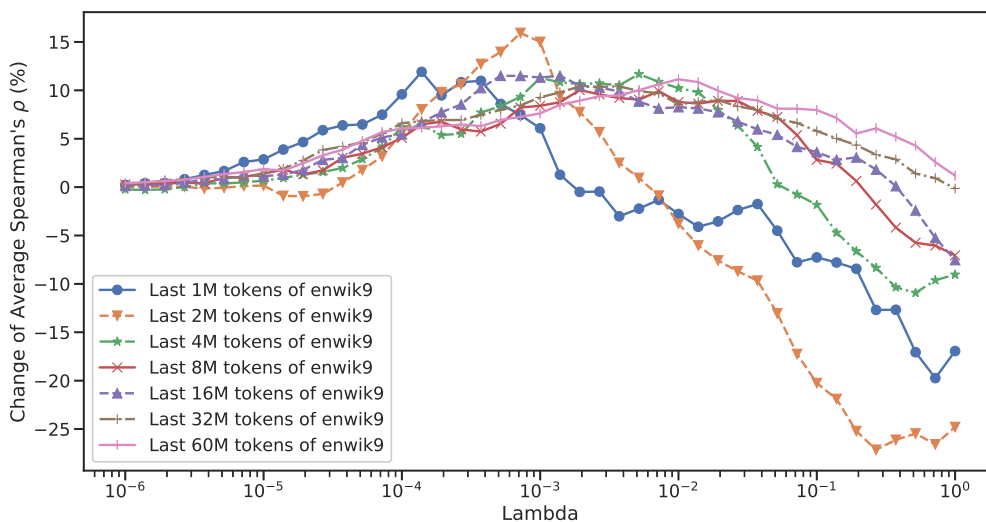


Figure 1: Influence of the λ -parameter on the average score (Spearman’s ρ) of all five used word similarity data sets (RG-65, WordSim-353, SimLex-999, MEN, RW) for different corpus sizes comparing 43 values of λ from 10^{-6} to 1.

For the experiments in the present paper, the best λ -values of the last 1M tokens were used for training word embeddings for the *first* 1M tokens of enwik9, the best λ -values of the last 2M tokens were used for training word embeddings for the *first* 2M tokens of enwik9 and so on. The best λ for the last 60 million tokens was used for word embeddings for both, the first 64 million tokens of enwik9 as well as the complete enwik9 corpus.

In the case of Maltese and Luxembourgish, the best λ -values were chosen conservatively according to the best values for the next smaller tested corpora, i.e., the best λ for the last 1M tokens of enwik9 for Maltese and that for the last 4M tokens of enwik9 for Luxembourgish.

4.2. Corpus Size

The results of the comparison of SGNS, PU-Learning for word embeddings, and the present approach SVD-PPMI $_{\lambda}$ for different corpus sizes are shown in figure 2.

It can be seen that SVD-PPMI $_{\lambda}$ outperforms the baseline approaches for most corpus sizes. Especially for the smaller sized corpora consisting of the first 1 and 2 mil-

lion tokens of enwik9, SVD-PPMI $_{\lambda}$ yields better results. For bigger corpora it performs very similarly to the PU-Learning approach. Taking a look at the performance on the RW (Rare Word) data set reveals that SVD-PPMI $_{\lambda}$ seems to work particularly well for representing rare words.

5. Related Work

5.1. Word Embeddings for Low-Resource Languages

Only few explicit approaches have been made to learn word embeddings for low-resource languages from scratch, i.e., without trying to project existing embeddings from other languages into the source language. Konovalov and Tumunbayarova (2018) investigated the training of word embeddings for the mongolic language of Buryat using the classical way of factorizing a PPMI matrix by truncated SVD. Jiang et al. (2018) proposed to learn word embeddings for low-resource languages from PPMI matrices by applying PU-Learning to overcome the sparsity of matrices caused by the lack of data. Neither of these methods used Dirichlet smoothing for PPMI.

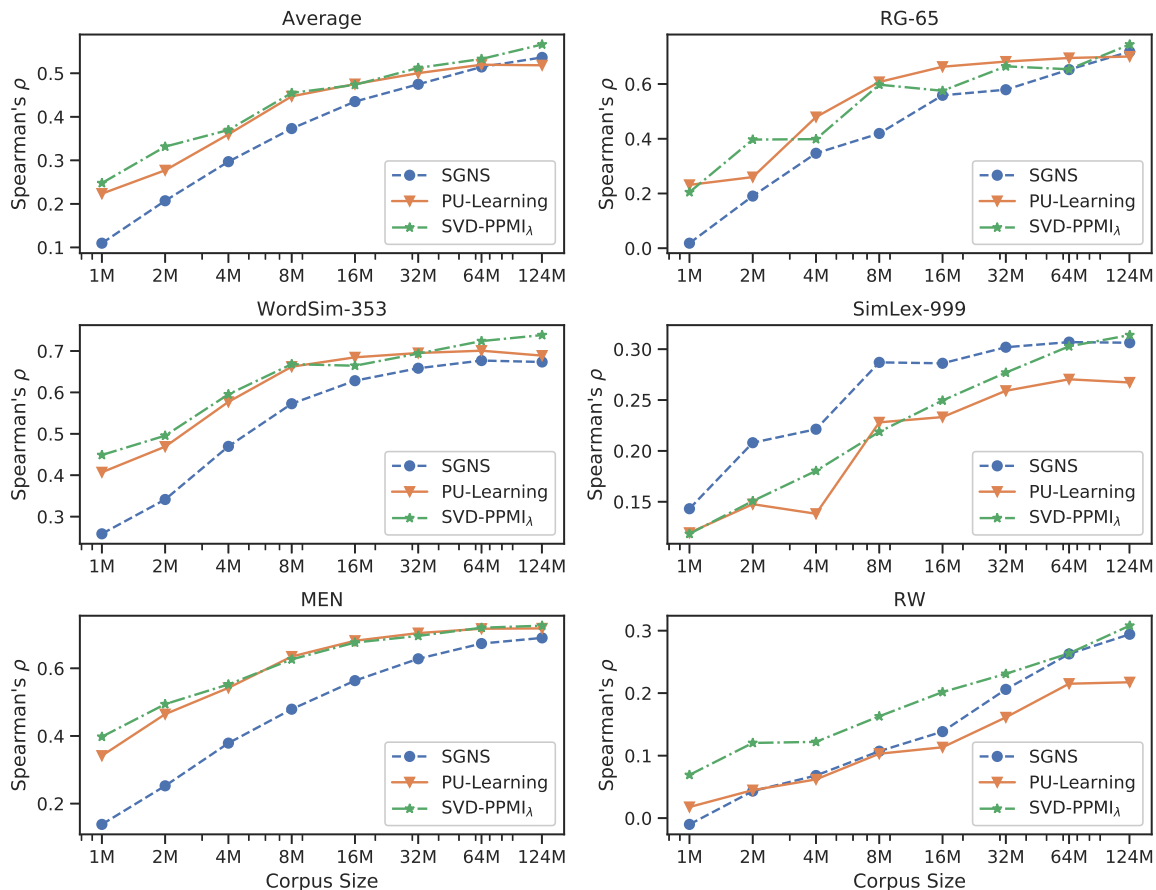


Figure 2: Performance of the methods on word similarity tasks as a function of the corpus size (for differently sized slices of enwik9). The top left graph shows the macro average of all five used word similarity data sets.

5.2. Smoothing the PMI Matrix

Levy et al. (2015) approach PMI's bias towards rare words by a method called *context distribution smoothing*, which is inspired by the choice of the noise distribution used by word2vec SGNS to generate negative samples. The idea is to substitute PMI by

$$\text{PMI}_\alpha(w, c) = \log \left(\frac{\hat{P}(w, c)}{\hat{P}(w)\hat{P}_\alpha(c)} \right) \quad (8)$$

where $\hat{P}_\alpha(c)$ is the smoothed context probability

$$\hat{P}_\alpha(c) = \frac{f(c)^\alpha}{\sum_{c' \in V_c} f(c')^\alpha}. \quad (9)$$

The effect of this smoothing technique is that, given c is rare, the probability of the context word will be higher than before, i.e., $\hat{P}_\alpha(c) > \hat{P}(c)$. This, in turn, reduces the PMI of co-occurrences with rare words.

Another possibility, proposed by Pantel and Lin (2002), is to multiply the PMI value by the following discount factor:

$$\delta_{w,c} = \frac{f(w, c)}{f(w, c) + 1} \cdot \frac{\min(f(w), f(c))}{\min(f(w), f(c)) + 1}. \quad (10)$$

This causes that, the less frequent one of w or c gets, the more the final weight $\text{PMI}_\delta(w, c)$ will be reduced. The left

factor in the equation causes a similar reduction if the co-occurrence count of the word pair is low. All in all, δ pushes the PMI values towards zero, more for rare words and less for frequent words or co-occurrences.

Turney and Pantel (2010) and Jurafsky and Martin (2018) mention the idea to use Dirichlet smoothing to weaken PMI's bias towards rare words. Turney and Littman (2003) used additive smoothing in the context of calculating association strength in order to avoid division by zero in their specific setting. But, to the best of our knowledge, the idea of factorizing Dirichlet-smoothed count matrices for obtaining word embeddings has not been carried out nor evaluated in previous work.

6. Conclusion

This work investigates classical PPMI embeddings with Dirichlet smoothing to correct its bias towards rare words. We show that classical PPMI based word embeddings can outperform machine-learning-based methods in a low-resource setting.

In a case study we demonstrated this on the low-resource languages Maltese and Luxembourgish. Further work should investigate its performance in domain-specific low-resource settings.

7. Acknowledgements

This work has been funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibilities for its content.

8. Bibliographical References

- Bruni, E., Tran, N. K., and Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Int. Res.*, 49(1):1–47, January.
- Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526.
- Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.
- Finkelstein, L., Gavrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppim, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on information systems*, 20(1):116–131.
- Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.
- Jiang, C., Yu, H.-F., Hsieh, C.-J., and Chang, K.-W. (2018). Learning word embeddings for low-resource languages by PU learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1024–1034, New Orleans, Louisiana, June. Association for Computational Linguistics.
- Jurafsky, D. and Martin, J. H. (2018). *Speech and language processing* (3rd edition draft).
- Kononov, V. P. and Tumunbayarova, Z. B. (2018). Learning word embeddings for low resource languages: The case of buryat. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018”*, pages 331–341.
- Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.
- Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.
- Luong, M.-T., Socher, R., and Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria, 08.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Pantel, P. and Lin, D. (2002). Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619. ACM.
- Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.
- Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.
- Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.
- Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.