Reinforcing the Topic of Embeddings with Theta Pure Dependence for Text Classification*

Ning Xing[1], **Yuexian Hou**[1], **Peng Zhang**[1], **Wenjie Li**[2], **Dawei Song**[1]
[1]School of Computer Science and Technology, Tianjin University, China
[2]Department of Computing, The Hong Kong Polytechnic University, Hong Kong
{xingning,yxhou,pzhang,dwsong}@tju.edu.cn,cswjli@comp.polyu.edu.hk

Abstract

For sentiment classification, it is often recognized that embedding based on distributional hypothesis is weak in capturing sentiment contrast–contrasting words may have similar local context. Based on broader context, we propose to incorporate Theta Pure Dependence (TPD) into the Paragraph Vector method to reinforce topical and sentimental information. TPD has a theoretical guarantee that the word dependency is pure, i.e., the dependence pattern has the integral meaning whose underlying distribution can not be conditionally factorized. Our method outperforms the state-of-the-art performance on text classification tasks.

1 Introduction

Word embeddings can be learned by training a neural probabilistic language model or a unified neural network architecture for various NLP tasks (Bengio et al., 2003; Collobert and Weston, 2008; Collobert et al., 2011). In global context-aware neural language model (Huang et al., 2012), the global context vector is a weighted average of all word embeddings of a single document/paragraph. After trained with all word embeddings belonging to the current paragraph, a resulting Paragraph Vector can be obtained. Actually, Le and Mikolov's Paragraph Vector (Le and Mikolov, 2014) is trained based on the log-linear neural language model (Mikolov et al., 2013a).

For text classification, using a straightforward extension of language model (e.g. Le and Mikolov's Paragraph Vector) is considered not to be sensible. Embeddings learned for text classification should be very different from that learned for language modeling. For example, language models often calculate the probability of a sentence, therefore *this is a good movie* and *this is a bad movie* may not be discriminated from each other. In sentiment analysis task, the semantic representation of words needs to tell word *good* from *bad*, even if the two words have the same local context. For this reason, the local dependency is insufficient to model topical or sentiment information. Fortunately, if we have the global context of *good* like *interesting* or *amazing*, the sentiment meaning of the embedding will be explicit. However, the training of log-linear neural language model is based on local word dependencies (e.g., the co-occurrence of the words in a local window). Thus, Paragraph Vector can not explicitly model the word dependencies for those words that do not frequently appear in a local window but are actually closely dependent on each other.

In this paper, our aim is to extend the Paragraph Vector with global context which can capture topical or sentiment information effectively. However, if one explicitly considers the dependency patterns that are beyond the local window level, there is a possibility that the noisy dependency patterns can be involved and modeled in the distributed representation methods. Moreover, there should be an unique and explicit topical meaning in the patterns to guarantee no ambiguity in the global context. Therefore, we need a dependency mining method that not only models the long range dependency patterns, but also provides a theoretical guarantee that the dependency patterns are pure. Here, the "pure" dependency pattern is an integral semantic meaning/concept that cannot be factorized into sub dependency patterns.

In the language of statistics, Conditional Pure Dependence (CPD) means that the underlying distribution of the dependency patterns cannot be factorized under certain conditions (e.g., priors, observed words, etc.). It has been proved that CPD is the high-level pure dependence in (Hou et al.,

*Corresponding authors: Yuexian Hou and Peng Zhang.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2551–2556,
Lisbon, Portugal, 17-21 September 2015.

2013). However, judging CPD is NP-hard (Chickering et al., 2004). Fortunately, Theta Pure Dependence (TPD) is the sufficient criteria of CPD and can be identified in O(N) time, where N is the number of words (Hou et al., 2013). This finding motivates us to adopt TPD as the global context. Moreover, compared with other conventional co-occurrence-based methods, such as the Apriori algorithm (Agrawal et al., 1993), TPD based on the Information Geometry (IG) framework has a solid theoretical interpretations in statistics to guarantee the dependence is pure.

2 Modeling Topic with TPD

Compared with local context, global context can usually capture the text topic more precisely. It is easy to get local context by a sliding window. We define the centered word as the *current word* and the other words in the window as *local context words*. Global context words are extracted from all the documents in the corpus and can be divided into two parts: a) the words in the current document but outside of the local context window; b) the words never appeared in the document but in the corpus. The following example shows the words mentioned above, and the topic (the scene of filming) is easily captured by TPD:

- TPD: **scene** camera acting movie
  Text: there [*is great atmosphere in the* **scene** *from the location , the*] lighting , the fog and such , but the camera should be slowly following the killer...

The bracket stands for the local context window, and the size of window is 5, i.e. there are five local context words (in italics) in both sides of the current word (in bold). Global context words are underlined in the example.

In order to model the topic explicitly, the dependence pattern should report one and only one topical meaning. TPD has a theoretical guarantee that the dependency has an integral meaning whose underlying distribution can not be conditionally factorized. Formally, given a set of binary random variables $\mathbb{X} = \{X_1, \ldots, X_n\}$, where X_i denotes the occurrence $(X_i = 1)$ or absence $(X_i = 0)$ of the i-th word. Then the n-order TPD over $\mathbb{X}$ can be defined as follows.

DEFINITION 1. (TPD): $\mathbb{X} = \{X_1, \ldots, X_n\}$ is of n-order Theta Pure Dependence (TPD), iff the n-order θ coordinate $\theta_{12\ldots n}$ is significantly different from zero. (Hou et al., 2013)

TPD can be effectively identified by an explicit statistical test procedure: Log Likelihood Ratio Test (LLRT) (Nakahara and Amari, 2002) for θ-coordinate of IG. (Hou et al., 2013)

Here, we introduce two negative examples to further emphasize the importance of utilizing TPD. Example 1: *can, with, of*. The joint distribution of this words combination can be unconditionally factorized directly, since the occurrence of any word does not necessarily imply the occurrence of others. Example 2: *London, Chelsea, Sherlock Holmes*. As we all know, both *Chelsea* and *Sherlock Holmes* are closely related to *London*. *Chelsea* and *Sherlock Holmes* are two relatively independent topics, i.e. they are conditional independent given *London*. Although the three phrases are unconditionally dependent, their joint distribution can be conditionally factorized. Thus the dependency in both two examples can not be pure.

To explain TPD and the characteristic "pure" intuitively, let us look at a typical example of TPD: *climate, conference, Copenhagen*. The co-occurrence of the three words implies an unseparable high-level semantic entity compared with the two negative examples, introduced above. In negative examples, the high frequency of words co-occurrence can be explained as some kind of "coincidence", because each of them or their pairwise combinations has a high frequency, independently. However, the co-occurrence of TPD words cannot be fully explained as the random coincidence of, e.g., the co-occurrence of *Copenhagen* and *conference* (which can be any other conferences in Copenhagen) and the occurrence of *climate*.

The word "pure" in Hou et al. (2013) means that the joint probability distribution of these words is significantly different from the product of lower-order joint distributions or marginal distributions, w.r.t all possible decompositions. More formally, it requires that the joint distribution cannot be factorized unconditionally (UPD) or conditionally (CPD) in the language of graphical model. Let $x_i \in \{0, 1\}$ denote the value of X_i. Let $p(x)$, $x = [x_1, x_2, \ldots, x_n]^T$, be the joint probability distribution over $\mathbb{X}$. Then the definitions of UPD and CPD are as follows:

DEFINITION 2. (UPD): $\mathbb{X} = \{X_1, \ldots, X_n\}$ is of n-order Unconditional Pure Dependence (UPD), iff it can NOT be unconditionally factorized, i.e., there does NOT exist a k-partition $\{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_k\}$ of $\mathbb{X}$, $k > 1$, such at $p(x) =$

$p(c_1) * p(c_2)...p(c_k)$, where $p(c_i)$, $i = 1, \ldots, k$, is the joint distribution over $\mathbb{C}_i$. (Hou et al., 2013)

DEFINITION 3. (CPD): $\mathbb{X} = \{X_1, \ldots, X_n\}$ is of n-order Conditional Pure Dependence (CPD), iff it can NOT be conditionally factorized, i.e., there does NOT exist $\mathbb{C}_0 \subset \mathbb{X}$ and a k-partition $\{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_k\}$ of $\mathbb{V} = \mathbb{X} - \mathbb{C}_0$, $k > 1$, such at $p(v|c_0) = p(c_1|c_0) * p(c_2|c_0)...p(c_k|c_0)$, where $p(v|c_0)$ is the conditional joint distribution over $\mathbb{V}$ given $\mathbb{C}_0$, and $p(c_i|c_0)$, $i = 1, 2, \ldots, k$, is the conditional joint distribution over $\mathbb{C}_i$ given $\mathbb{C}_0$. In case that $\mathbb{C}_0$ is an empty set, we define $p(c_0) = 1$. (Hou et al., 2013)

Actually, CPD is stricter than UPD, and the dependence which just satisfies UPD is not pure enough to model the global context. Therefore, "pure" in our paper refers to the characteristic of CPD. However judging CPD is NP-hard. It is proved that a significant nonzero n-order θ parameter (TPD) entails the n-order CPD/UPD in Hou et al. (2013). The highest-order coordinate parameter in IG is a proper metric for the purity (i.e., the unique semantics) of high-order dependence. A pattern is TPD, iff the n-order θ coordinate $\theta_{12...n}$ is significantly different from zero. Moreover, The Log Likelihood Ratio Test implemented in the mixed coordinates can test whether $\theta_{12...n}$ is significantly different from zero.

Contrasting to TPD, the semantic coupling among the associations in the two negative examples is much weaker. In conclusion, *can, with, of* cannot give an explicit topic and *London, Chelsea, Sherlock Holmes* includes at least two topics. the co-occurrence of words in TPD (e.g. *climate, conference, Copenhagen*) implies an un-separable (pure) high-level semantic entity. A sufficient and unbroken meaning of dependence can not only supply the context but also avoid the ambiguity (or noise) in global context. Therefore, the meaning of pure is important in such a global context modeling method.

3 Global PV-DBOW and Dependence Vectors

A version of Paragraph Vector in Le and Mikolov (2014) PV-DBOW is extended with TPD to a new model: Global PV-DBOW (Glo-PV-DBOW). TPD has been extracted from the corpus before training. Given a sequence of training words w_1, w_2, w_3, ..., w_T and the global context glo_t of w_t, the objective of Glo-PV-DBOW is to maximize the average log probability:

$$L = \frac{1}{T}\sum_{t=1}^{T}\Big[\sum_{-c\le j\le c, j\neq 0} \log p(w_t|w_{t+j}) + \log p(w_t|glo_t) + \log p(w_t|doc_t)\Big] \quad (1)$$

where c is the local context window size. The indicator of the document that the current word w_t belongs to is denoted by doc_t. Further, we define $p(w_t|glo_t)$ in equation (2):

$$p(w_t|glo_t) = \prod_{a}^{A}\Big[p(w_t|dep_t^a)p(w_t|w_1^a, w_2^a, \ldots, w_N^a)\Big] \quad (2)$$

The indicator of the a-th w_t's TPD pattern is denoted as dep_t^a and can be trained to be a distributed representation of TPD: dependence vector $v_{dep_t^a}$. This (N+1)-order TPD consists of N+1 words: w_1^a, $w_2^a...w_N^a$ and w_t. The energy function of w_t and $w_i = (w_{t+j}, doc_t, dep_t^a)$ is uniform as follows:

$$E(w_t, w_i) = -{v_{w_t}}^T v_{w_i} \quad (3)$$

We define the energy function of TPD words:

$$E(w_t, w_1^a, w_2^a, \ldots w_N^a) = -\frac{1}{N}\sum_{n=1}^{N}{v_{w_t}}^T v_{w_n^a} \quad (4)$$

The resulting predictive distributions are given by

$$p(w_t|w_i) = \frac{exp({v_{w_t}}^T v_{w_i})}{\sum_{m=1}^{W} exp({v_{w_m}}^T v_{w_i})} \quad (5)$$

$$p(w_t|w_1^a, w_2^a, \ldots, w_N^a) = \frac{exp(\frac{1}{N}\sum_{n=1}^{N}{v_{w_t}}^T v_{w_n^a})}{\sum_{m=1}^{W} exp(\frac{1}{N}\sum_{n=1}^{N}{v_{w_m}}^T v_{w_n^a})} \quad (6)$$

Hierarchical softmax (Morin and Bengio, 2005) is adopted to reduce the cost of computation. The binary tree is specified with a Huffman tree, and the Huffman code of pseudo words m_i in w_t's Huffman path is denoted as x_{m_i}. For more about hierarchical softmax we used, please refer to (Mikolov et al., 2013b). Using stochastic gradient descent (SGD), distributed representations of the word, dependence and document have been trained. The update procedure of $v_{w_i} = (v_{w_{t+j}}, v_{doc_t}, v_{dep_t^a})$ is as same as the procedure described in (Mikolov et al., 2013b). Thus, the pseudo code for training TPD words is listed individually:

SGD FOR TRAINING THE TPD WORDS

$v_{w_t} \leftarrow current\_word$
$v_{w^a_{ave}} \leftarrow \frac{1}{N}\sum_{n=1}^{N} v_{w^a_n}$
$err \leftarrow 0$
for $\forall m_i$
    do $g \leftarrow (1 - x_{m_i} - \sigma({v_{m_i}}^T v_{w^a_{ave}})) * \alpha$
        $err+ = g * \frac{1}{N} * v_{m_i}$
        $m_i+ = g * v_{w^a_{ave}}$
for $n \leftarrow 1$ **to** N
    do $v_{w^a_n}+ = err$

4 Experiments

Apriori (not a pure dependency method) is contrastively adopted to implement Glo-PV-DBOW. Glo-PV-DBOW-TPD and Glo-PV-DBOW-Apri are all evaluated in two text classification tasks: sentiment analysis and topic discovery. The suffix (e.g., -2, -5) of our global method name denotes the order of dependency (the number of words in a dependence pattern). The order of dependency is changed because we want to show the superiority of the high-order TPD. The high-order TPD provides the more rich and explicit global context than the lower-order one since the high-order TPD cannot be reduced to the random coincidence of lower-order dependencies.

We cross-validate the hyperparameters and set the local context window size as 10, the dimension of embeddings as 100. In sentiment analysis task, Apriori's minimum support and TPD's theta_0 is respectively set as 0.004 and 1.4. While in topic discovery task, Apriori's minimum support and TPD's theta_0 is around 0.020 and 2.0 respectively. Since the classification accuracy of the approaches compared is a single result, we do not include any results for test of significance in our method and only report the average accuracy.

4.1 Sentiment Analysis on Movie Reviews

The binary sentiment classification on the IMDB dataset proposed by (Maas et al., 2011) is conducted. Results in Fig.1 show that global methods' performance is more stable than PV-DBOW's. Moreover, TPD works much better than Apriori, especially in the high-order dependence. Note that TPD-5 works better than TPD-2, while Apri-5 works worse than Apri-2. It can be explained that the Apriori algorithm is short of an explicit statistical test procedure to guarantee the pure dependence. Therefore, the Apriori algorithm is not suitable for generating the high-order dependence.

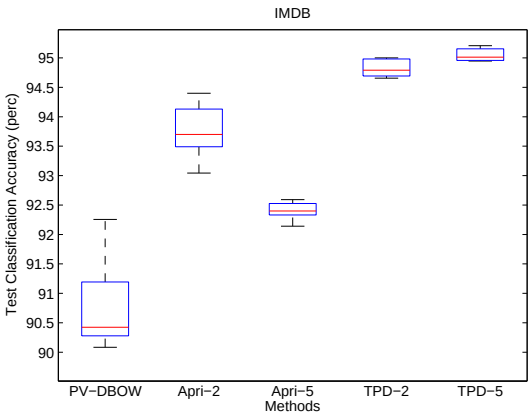

Figure 1: Box plot of classification accuracy over a local method (PV-DBOW) and 4 global methods (Apri-2/5, TPD-2/5).

Instead, the high-order TPD can provide the rich and explicit global context for the model. Meanwhile, it is verified that our method is good at capturing sentiment contrast.

Table 1 shows that Glo-PV-DBOW with 5-order TPD achieves the state-of-the-art performance. A promising result is an improvement of more than 2% over result published in Le and Mikolov (2014). Note that the algorithm process of Paragraph Vector (Le and Mikolov, 2014) is much more complex than PV-DBOW's. Paragraph Vector includes an extra inference stage. In addition, Paragraph Vector's document vector is a combination of two vectors: one learned by PV-DBOW and the other learned by Distributed Memory Model of Paragraph Vectors (PV-DM) (Le and Mikolov, 2014). The combined document vector has 800 dimensions, while all vectors in our experiments only have 100 dimensions.

4.2 Topic Discovery on News

The 20 Newsgroups dataset is a collection of approximately 20,000 newsgroup documents, partitioned across 20 different newsgroups. We follow (Crammer et al., 2012) to create binary problems from the dataset by creating binary decision problems of choosing between two similar groups. Therefore, the dataset is split into two sub-datasets as follows: *comp: comp.sys.ibm.pc.hardware vs. comp.sys.mac.hardware* and *sci: sci.electronics vs. sci.med*. Similarly, 1800 examples balanced between the two labels were selected for each problem.

The classification accuracy on each sub-dataset

Table 1: The performance of our method compared with other approaches on the IMDB dataset.

Model	Accuracy rate
BoW (bnc) (Maas et al., 2011)	87.80%
Full+Unlabeled+BoW (Maas et al., 2011)	88.89%
WRRBM (Dahl et al., 2012)	87.42%
WRRBM + BoW (bnc) (Dahl et al., 2012)	89.23%
SVM-bi (Wang and Manning, 2012)	89.16%
NBSVM-bi (Wang and Manning, 2012)	91.22%
PV-DBOW (Le and Mikolov, 2014)	90.79%
Paragraph Vector (Le and Mikolov, 2014)	92.58%
Sentence Vector + RNN-LM + NB-LM (Mesnil et al., 2014)	92.57%
mvCNN**o&w** (Johnson and Zhang, 2015)	93.34%
Glo-PV-DBOW-Apri-2	93.76%
Glo-PV-DBOW-Apri-5	92.41%
Glo-PV-DBOW-TPD-2	94.83%
Glo-PV-DBOW-TPD-5	**95.05%**

Table 2: The performance of our method compared to other approaches on 20 Newsgroup.

Model	Comp	Sci
Confidence-weighted (Crammer et al., 2012)	94.39%	97.56%
PV-DBOW (Le and Mikolov, 2014)	92.60%	98.02%
Glo-PV-DBOW-Apri-2	94.56%	98.42%
Glo-PV-DBOW-Apri-5	94.43%	98.13%
Glo-PV-DBOW-TPD-2	94.59%	**99.20%**
Glo-PV-DBOW-TPD-5	**95.47%**	98.74%

is recorded in Table 2. Compared with Confidence-weighted (Crammer et al., 2012) and PV-DBOW (Le and Mikolov, 2014), our extended models achieve the highest accuracy on each subdataset. Moreover, TPD as a pure dependence works better than Apriori when they provide the global context for our model. The topical information is effectively reinforced in embeddings by incorporating TPD.

4.3 Analysis on Word Embeddings

The cosine similarity of each word pair in 20 Newsgroups is computed. We list four center words and their nearest neighbors in PV-DBOW and Glo-PV-DBOW groups respectively. The rankings are labeled in front of neighbor words, and some notable neighbor words are in bold.

From Table 3, we can see that the statistical information of corpus like words co-occurrence can be mined by TPD. Therefore, the Glo-PV-DBOW's embeddings are context-aware and it can help a lot for classification tasks. The top 40 nearest neighbors of *ibm* are investigated, and we find *macintosh* and *mac* appeared in the PV-DBOW group but not in the Glo-PV-DBOW group. In

Table 3: Nearest neighbors of words ranking list based on cosine similarity.

Center word	PV-DBOW	Glo-PV-DBOW
ibm	1:aix	1:aix
	2:pc	2:pc
	...	3:pc's
	23:**macintosh**	4:**austin**
	34:**mac**	5:**workstations**
mac	1:macintosh	1:macintosh
	2:quicktime	2:**apple's**
	3:portable	3:quicktime
	4:utilities	4:**apple**
	5:**macs**	5:**macs**
486	1:386	1:386
	2:486dx	2:**cpu**
	3:33mhz	3:486dx
	4:486dx2	4:486dx2
	5:**cpu**	5:33mhz
Kingston	1:aix	1:aix
	2:mike	2:**ibm**
	3:sharks	3:jones
	4:jones	4:sharks
	5:**ibm**	5:mike

the corpus, the topic of documents is either *ibm* or *mac*. If we perform a classification task on "*ibm* versus *mac*", it will be hard to classify in the PV-DBOW group. That is because PV-DBOW tends to regard *ibm* and *mac* both as computers. However, the two different computer brands are distinguished in Glo-PV-DBOW. Further, *ibm* and *mac* co-occur rarely in one document, and the statistical information is noted by TPD.

5 Conclusion

This paper proposes to incorporate Theta Pure Dependence into Paragraph Vector to capture more topical and sentimental information in the context. The extended model is applied to a sentiment classification task and a topical detection task. Our accuracy outperforms the state-of-the-art result on the movie and news datasets. The approach can be improved further to fully leverage the un-factorized sense of high-order Theta Pure Dependence. In future, we will explore the applications of dependence distributed representation.

Acknowledgments

This work is funded in part by the Chinese 863 Program (grant No. 2015AA015403), the Key NSF Project of Chinese Tianjin (grant No. 15JCZDJC31100), the Chinese 973 Program (grant No. 2013CB329304 and 2014CB744604), the Chinese NSF Project (grant No. 61272291, 61402324 and 61272265).

References

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155.

David Maxwell Chickering, David Heckerman, and Christopher Meek. 2004. Large-sample learning of bayesian networks is np-hard. *The Journal of Machine Learning Research*, 5:1287–1330.

Ronan Collobert and Jason Weston. 2008. A unified aagrawal1993miningrchitecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Koby Crammer, Mark Dredze, and Fernando Pereira. 2012. Confidence-weighted linear classification for text categorization. *The Journal of Machine Learning Research*, 13(1):1891–1926.

George E Dahl, Ryan P Adams, and Hugo Larochelle. 2012. Training restricted boltzmann machines on word observations. *Proceedings of International Conference on Machine Learning*.

Yuexian Hou, Xiaozhao Zhao, Dawei Song, and Wenjie Li. 2013. Mining pure high-order word associations via information geometry for information retrieval. *ACM Transactions on Information Systems (TOIS)*, 31(3):12.

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

Rie Johnson and Tong Zhang. 2015. Semi-supervised learning with multi-view embedding: Theory and application with convolutional neural networks. *arXiv preprint arXiv:1504.01255*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 142–150. Association for Computational Linguistics.

Grégoire Mesnil, Marc'Aurelio Ranzato, Tomas Mikolov, and Yoshua Bengio. 2014. Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. *arXiv preprint arXiv:1412.5335*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *ICLR Workshop*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Frederic Morin and Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In *Proceedings of the international workshop on artificial intelligence and statistics*, pages 246–252. Citeseer.

Hiroyuki Nakahara and Shun-ichi Amari. 2002. Information-geometric measure for neural spikes. *Neural Computation*, 14(10):2269–2316.

Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics.