ANTUSD: A Large Chinese Sentiment Dictionary

Shih-Ming Wang and Lun-Wei Ku
Academia Sinica
128 Academia Road, Section 2, Nankang,
Taipei 11529, Taiwan
E-mail: ipod825@gmail.com, lwku@iis.sinica.edu.tw

Abstract

This paper introduces the augmented NTU sentiment dictionary, abbreviated as ANTUSD, which is constructed by collecting sentiment stats of words in several sentiment annotation work. A total of 26,021 words were collected in ANTUSD. For each word, the CopeOpi numerical sentiment score and the number of positive annotation, neutral annotation, negative annotation, non-opinionated annotation, and not-a-word annotation are provided. Words and their sentiment information in ANTUSD have been linked to the Chinese ontology E-HowNet to provide rich semantic information. We demonstrate the usage of ANTUSD in polarity classification of words, and the results show that a superior f-score 98.21 is achieved, which supports the usefulness of the ANTUSD. ANTUSD can be freely obtained through application from NLPSA lab, Academia Sinica: http://academiasinicanlplab.github.io/

Keywords: opinion dictionary, sentiment dictionary, NTUSD, ANTUSD

1. Introduction

Sentiment analysis and opinion mining has been an important sub-discipline in natural language processing, information retrieval and machine learning. Related techniques have many applications, such as product/travel/movie recommendation, automatic opinion poll, melancholia detection, etc., so not only researchers try to build cutting edge systems to find sentiment information, but industries are searching for good solutions. An especially useful resource for both the research and industrial communities is a big and complete sentiment dictionary which can serve as the material for multiple purposes, e.g., markers of opinions or machine learning features. However, we are lacking this kind of resources for the Chinese language. In this paper we introduce the augmented NTU sentiment dictionary, abbreviated as ANTUSD, which is constructed by collecting sentiment stats of words in several sentiment annotation work. It collects 27,221 words. For each word, the CopeOpi numerical sentiment score and the number of positive annotation, neutral annotation, negative annotation, non-opinionated annotation, not-a-word annotation are provided. To extend the usage of ANTUSD, it has been connected to a large Chinese ontology, E-HowNet, to provide rich semantic information to the dictionary entries. In the last part of this paper, we show that using ANTUSD in sentiment word detection and word polarity classification both achieve good results.

2. Related Materials

ANTUSD was built mainly by collecting manual annotation of words in the process of building several sentiment corpora in a long period of time from year 2006 to year 2010. Therefore, before we describe the construction of ANTUSD, materials involved in this process are first introduced: NTUSD, NTCIR MOAT task dataset, Chinese Opinion Treebank (L.-W. Ku, Huang, & Chen, 2010), AciBiMA (T.-H. Huang, Ku, & Chen, 2010; T.-H. K. Huang, Chen, & Kong, 2015), CopeOpi and E-HowNet, where NTUSD is a sentiment dictionary, NTCIR MOAT task dataset and Chinese opinion Treebank are two manual labeled opinion sentence datasets, CopeOpi is a Chinese opinion scoring system, and E-HowNet is a Chinese knowledge ontology, described as follows:

NTUSD NTUSD (L.-W. Ku, Liang, & Chen, 2006) is the prototype sentiment dictionary of ANTUSD. It provides a total of 11,088 sentiment words containing 2,812 positive words and 8,276 negative words. NTUSD was published in year 2006 and has been downloaded more than 300 times and widely adopted by researchers in sentiment analysis area. NTUSD provides useful polarity information which can serve as seeds to learn sentiment of other words, sentences and even documents. However, it provided no detail information for the polarity strength of these words. ANTUSD is an enhanced version from this aspect. It not only covers more words than NTUSD, and also gives each word a numerical sentiment score and the numbers of labels so far annotated in the process of annotating several sentiment resources. They could be used to estimate the sentiment as well as the strength of each word.

NTCIR MOAT Task Dataset The sentiment labels of each word in ANTUSD were collected from two datasets: NTCIR MOAT Task Dataset and Chinese Opinion Treebank. On these two datasets, sentence-level sentiment labels were both annotated. Three NTCIR[1] tasks were related to opinion analysis, including NTCIR-6 pilot opinion task (OAT, 2006/2007), NTCIR-7 MOAT (MOAT1, 2007/2008) and NTCIR-8 (MOAT 2, 2009/2010), traditional Chinese side. In these NTCIR tasks, each sentences were labeled by three annotators. When annotators read the sentences, they also labeled the sentiment words they found. Same to sentences, words were labeled as positive, neutral, and negative.

Chinese Opinion Treebank To further incorporate the syntactic information, i.e., parse trees, into sentiment

[1] http://research.nii.ac.jp/ntcir/index-en.html

analysis, Ku has developed the Chinese Opinion Treebank (L.-W. Ku et al., 2010). Chinese Opinion Treebank is developed by labeling on the sentences in Chinese Treebank 5.1[2]. First all the Chinese sentences were labeled as opinionated and non-opinionated. Then more detail sentiment information was labeled on opinionated sentences. When labeling these opinionated sentences, annotators also labeled the sentiment words they found as positive, neutral and negative. Note that accumulated number of labels collected in the annotation process is related to the word frequency, as annotators will only label sentiment words when they read them in documents or sentences.

ACIBiMA ACIBiMA is a Chinese word morphological structure corpus developed by Huang (T.-H. K. Huang et al., 2015). ACIBiMA now contains more than 10,000 Chinese words and their morphological types. It was a continuous work of building the Chinese Morphology Dataset (T.-H. Huang et al., 2010) and has been utilized to test the connection between the Chinese morphological structure and the Chinese sentiment (L.-W. Ku, Huang, & Chen, 2009). The Chinese word sentiment in this dataset were labeled by the annotators as positive, neutral, negative, non-opinionated, and not-a-word. As the words for labeling were selected randomly from a large automatically Chinese-word-segmented dataset, the label not-a-word is used to note those segmented incorrectly by the word segmentation system. Though the label not-a-word is not related to sentiment, it is included in ANTUSD as most Chinese text analysis tasks involve word segmentation and including words of the not-a-word type may provide some additional clues to ignore the word candidates from segmentation errors.

CopeOpi CopeOpi is a Chinese opinion-analysis system proposed in 2009 (L. W. Ku, Ho, & Chen, 2009). It determines the sentiment by accumulating the sentiment of the composite components, where words are determined by the composite characters, sentences by the composite words, and documents by the composite sentences. When determining the sentiment of words, CopeOpi calculates the degrees of the positive and negative polarity of characters by the observation probability in seed words of these two types, and uses these two values and a scoring function to reports the final sentiment score, ranged from -1 to 1, for each Chinese word. The CopeOpi word sentiment scores for all entries in ANTUSD were calculated and included.

E-HowNet Extended-HowNet (short as E-HowNet, CKIP 2009) is a frame-based entity-relation model extended from HowNet (Dong & Dong, 2006) to define lexical senses (concepts). It is a Chinese ontology of the lexical semantic representation. E-HowNet contains more than 80,000 Chinese words. However, as words in ANTUSD were collected from various sources, not all of them were included in E-HowNet. Therefore, through the integration with E-HowNet, E-HowNet can provide both the sentiment information and the lexical semantic information for 12,995 Chinese words, which cover 47.74 percentage of ANTUSD. Table 1 shows more detailed stats.

Data Set	**POS**	**NEU**	**NEG**
ANTUSD	9,382	16	11,224
ANTUSD in E-HowNet with labels	3,881	11	4,569
Coverage	41.37%	68.75%	40.71%
Data Set	**NONOP**	**NOT**	**Total**
ANTUSD	5,415	612	27,221
ANTUSD in E-HowNet with labels	3,872	506	12,995
Coverage	71.51%	82.68%	47.74%

Table 1: Number of Words with Gold Sentiment Labels.

	Granularity	Collected Label	Context
NTUSD	Word	Gold	Independent
NTCIR MOAT	Sentence	All	Dependent
Chinese Opinion Treebank	Sentence	All	Dependent
ACBiMA	Word	Gold	Independent

Table 2. Annotation Scheme of Related Corpora (Here "Gold" denotes that only one gold label determined by labels from all annotators for each word was collected, while "All" denotes that labels from annotators for each word were all collected.)

W	Score	Pos	Neu	Neg	Non	Not
勝利 (victory)	0.60	6	0	0	0	0
失敗 (failure)	-0.85	0	0	5	0	0
不致 (not to)	-0.05	0	4	2	0	0
大上 (big up)	0	0	0	0	0	5

Table 3. Example Entries in ANTUSD

3. Building ANTUSD

As mentioned in the previous section, the process of building ANTUSD is also the process of building the sentiment corpora NTCIR MOAT datasets, Chinese Opinion Treebank and the ACBiMA. ANTUSD includes all sentiment words originally in NTUSD, and then collects sentiment words annotated by the annotators working for NTCIR MOAT datasets, Chinese Opinion Treebank, and the ACBiMA. However, as these sentiment corpora were for exploring different research problems, their annotation scheme were also different. NTUSD is a sentiment word dictionary, and the sentiment annotation is context-free. Three annotators will label one candidate word and its gold sentiment is determined by majority. NTCIR MOAT

[2] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T01

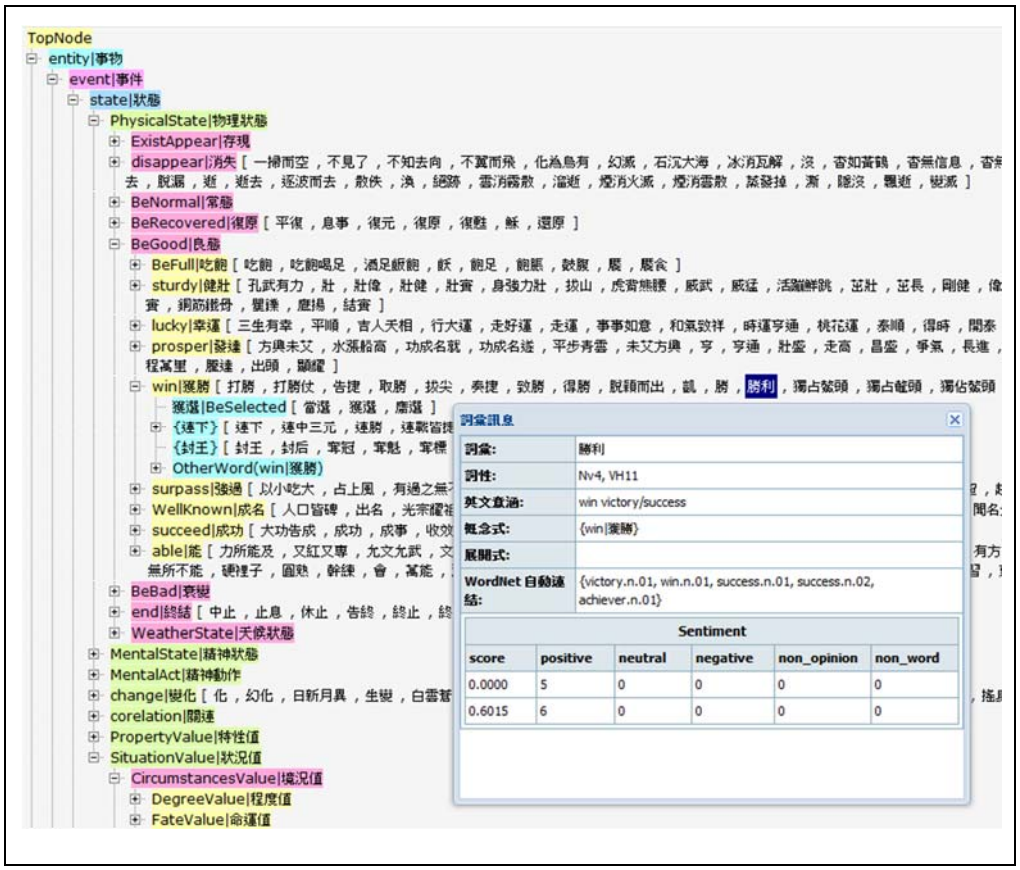

Figure 1. An E-HowNet Example Entry 勝利 (victory) with ANTUSD Information.

datasets and the Chinese Opinion Treebank are both sentence-level sentiment corpora. Same here, three annotators will label one candidate sentence and its gold sentiment is determined by majority. However, the annotators labeled the sentiment words they found at the time they read the candidate sentence, but we cannot guarantee all annotators will label the same words in one sentence. Therefore, all labels of words were collected from these two corpora, and as the word labels may depend on the content of the current sentences, these labels were context-dependent. ACBiMA has yet another labeling scheme. As the annotators familiar with Chinese morphology was difficult to find, only two annotators labels for each candidate Chinese word. When there was an inconsistency, they discussed and reached a conclusion as the gold label. As ACBiMA is also a dictionary and there is no context when annotators generated their sentiment labels, these labels are also context-free. A total of six fields are provided by ANTUSD for each word: the CopeOpi numerical sentiment score (Score), the number of positive annotation (*Pos*), neutral annotation (*Neu*), negative annotation (*Neg*), non-opinionated annotation (*Non*), and not-a-word annotation (*Not*). The annotation scheme of all related corpora are listed in Table 2 and Table 3 shows some example entries of ANTUSD. Figure 1 illustrates what ANTUSD information users can find from E-HowNet after the integration.

4. Using ANTUSD

As the first attempt towards applying ANTUSD, we performed sentiment analysis experiments on the words in ANTUSD. First, we derived a subset of words of ANTUSD and assign a unique label to each word of the subset according to its accumulative numbers of different labels in ANTUSD. We then performed three sentiment analysis tasks on the derived dataset by building SVM classifiers using three features: the CopeOpi score, the synonym-set index (SSI), and the word embedding.

4.1 Experimental Dataset Construction

As mentioned, ANTUSD provides numbers of manually-labeled sentiment instead of a single label for each word. For experiment, we need to assign a reliable sentiment label to each word. This procedure is described in Figure 2. Table 1 summarizes the numbers of words with gold labels. We then dropped possibly non-regular words (labeled as **NOT**). Words labeled as **NEU** (neural words) were also dropped since there are only 16 such words. As a result, we only use words with **POS**, **NEG**, **NONOP** labels in our experiments.

4.2 Word Features

As ANTUSD has been integrated with E-HowNet, the sentiment information from ANTUSD, e.g., the CopeOpi score, and the semantic information from E-HowNet, e.g., the synonym set information have the opportunity to serve as features together for sentiment analysis. To know how these features can help in sentiment analysis, we also implemented word embedding, the dense vector representation of words, as baselines for comparison.

1. Label the word **NOT** if Count(*Not)* >0
2. Label the word **NONOP** if Count(*Non)* >0
3. Label the word **POS** if Count(*Pos)* >0 and Count(*Neg*) =0
4. Label the word **NEG** if Count(*Neg*) >0 and Count(*Pos*) =0
5. Label the word **NEU** if Count(*Pos*) =0 and Count(*Neg*) =0 and Count(*Neu*) >0
6. Drop all not labeled words

Figure 2. Steps of Gold Sentiment Label Assignment.

CopeOpi Score The sentiment score of each Chinese word is determined by the sentiment scores of the component characters and the morphological type to form the word. In our experiments, CopeOpi scores (COP) in ANTUSD is used as one of our features.

Synonym-Set Index (SSI) E-HowNet defines a concept topology, represented as a directed graph. There are several general concept nodes connected to the root node, each of which connects to several more specific concept nodes as their children nodes. Each word in the topology is connected to at least a concept node as its parent node. Leaf node words of the same parent node could be regarded as synonyms. To index all internal nodes (concepts) to represent each synonym set, we encoded each word with a binary-coded vector. Each dimension of this word vector is set to 1 if the current word belongs to the corresponding synonym set, i.e., has a certain parent node. Note that as ANTUSD contains words that are not in E-HowNet, SSI is only accessible for the words in the intersection of ANTUSD and E-HowNet.

Word Embedding The previous two features can only provide partial information. CopeOpi considers only sentiment information. E-HowNet is manually labeled and hence may have the coverage problem. Hence, we consider word embedding as the third feature to provide the semantic of words. There has been a surge of research on representing words as dense vectors (Collobert & Weston, 2008; Fan, Chang, Hsieh, Wang, & Lin, 2008; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Mnih & Hinton, 2009). The concept of word embedding is based on the distributional hypothesis (Harris, 1954), which states that words in similar context should have similar meanings.

In our experiments, we trained word vectors of dimension 500 with the word2vec software package [3] from the corpus LDC2009T14 (C.-R. Huang, 2009). Since Chinese words are composed of characters, we considered representing each word with its word vector (WV) or the summation of its component character vectors (CV). Word vectors are trained with the word-tokenized corpus, and character vectors are trained by the character-tokenized corpus by further segmenting words into single characters. While all characters in the intersection of ANTUSD and E-HowNet can be found in LDC2009T14, some words in our experiment cannot be found in the corpus and hence lack corresponding word vectors. For such words, we set their WV to the 0 vector.

[3] https://code.google.com/p/word2vec

4.3 Experiment Setting

We performed three sentiment analysis tasks: opinion extraction, polarity classification and the combined task. Opinion extraction aims to separate opinion words (**POS**, **NEG**) from non-opinion words (**NONOP**). Polarity classification, on the other hand, classifies only opinion words by their polarities (**POS**, **NEG**). As for the combined task, we trained a three-label (**POS**, **NEG**, and **NONOP**) classifier. The performance of opinion extraction and polarity classification were both evaluated by the conventional f_1 score, while the combined task was evaluated by the following f-score:

$$\mathrm{p} = \frac{correct(opinion) \cap correct(polarity)}{proposed(opinion)},$$
$$\mathrm{R} = \frac{correct(opinion) \cap correct(polarity)}{gold(opinion)}, \quad (1)$$
$$f-score = \frac{2 \cdot P \cdot R}{P + R}.$$

All our experiments were conducted on 12,995 words shown in Table 1. Linear SVM (Fan et al., 2008) classifiers were adopted. For each task, we performed a 10-fold cross-validation, selected hyper-parameters leading to the highest average f score, and reported the average (over the 10 folds) precision, recall, and f_1/f-score under the best parameter.

4.4 Results and Discussion

The experiment results are shown as in Table 4, 5 and 6.
For opinion extraction, both COP and SSI show a much lower precision than WV and CV. Moreover, combining COP or SSI with WV or CV overall does not lead to a significant improvement. Since performance on opinion extraction may depend more on how much semantic information is captured in the feature vectors. It's not surprising that COP, a sentiment-oriented score, performs poorer in this task. The poorer performance of SSI might be due to the reason that the classification in the ontology provides little information of how words will be used in real context. In fact, in E-HowNet, there are many synonym sets containing only one word, making corresponding dimensions of the word vector very uninformative. Note that WV outperforms CV, this indicates that replacing word vectors with summation of component character vectors might lead to a less precise semantic representation.

For polarity classification, COP leads to a significant better result, which reflects its sentiment-oriented nature. However, combining other features with COP still leads to significant improvement, indicating that adding semantic information helps for polarity classification. Among the other features, WV is still the most informative feature. However, it does not dominate SSI, indicating the possibility for fine tuning the word embedding with prior knowledge (SSI in our case) as in (Faruqui et al., 2014).

For the combined task, COP outperforms all other combination of features since both the numerator of precision and the numerator of recall in Equation 1 are boosted up by COP's better polarity classification ability while only the denominator of precision is affected by COP's worse opinion extraction ability. All combined

features significantly outperforms their component features, indicating each feature is complementary for one another. Particularly, the reason that combining CV with WV outperforms WV is due to WV's smaller coverage (using 0 vectors for unseen words), which indicates that using CV might ease the pain of encountering new words.

Feature(s)	Precision	Recall	f-score
COP	0.6858	1	0.8136
SSI	0.6927	0.993	0.8161
WV	0.7844	0.936	0.8535
CV	0.765	0.9193	0.8350
COP+SSI	0.7395	0.9138	0.8175***
COP+WV	0.7849	0.9334	0.8527***
COP+CV	0.7639	0.9166	0.8332*
SSI+WV	0.7887	0.9367	0.8563**
SSI+CV	0.7724	0.9196	0.8396
WV+CV	0.8075	0.9213	0.8606

Table 4: Results of opinion extraction.
(In the last column, No tailed * sign means that a combined feature is significantly different from both its component feature at significance level 2.5%. A tailed * sign denotes that the combined feature is not significantly different from one (or both) of its component feature at significance level 2.5%. Tailed ** at 5%. Tailed *** at 7.5%.)

Feature(s)	**POS** f_1	**NEG** f_1	Average f_1
COP	0.9728	0.9757	0.9742
SSI	0.7918	0.8424	0.8171
WV	0.8702	0.8945	0.8824
CV	0.829	0.8509	0.8399
COP+SSI	0.9788	0.9815	0.9801
COP+WV	0.9806	0.9836	0.9821
COP+CV	0.9674	0.9721	0.9698*
SSI+WV	0.8979	0.9154	0.9066
SSI+CV	0.8678	0.8858	0.8768
WV+CV	0.8994	0.9157	0.9076

Table 5: Results of polarity classification. The meaning of * sign is the same as in Table 6.

Feature(s)	Precision	Recall	*f-score*
COP	0.9124	0.9272	0.9197
SSI	0.7064	0.6786	0.6922
WV	0.7371	0.7669	0.7517
CV	0.6887	0.7212	0.7045
COP+SSI	0.864	0.9446	0.9025
COP+WV	0.8497	0.902	0.875
COP+CV	0.8396	0.8686	0.8538
SSI+WV	0.7641	0.7955	0.7794
SSI+CV	0.7321	0.755	0.7434
WV+CV	0.7636	0.8129	0.7874

Table 6: Results of the combined task.

5. Conclusion and Future Work

In this paper, we have constructed so far the largest Chinese sentiment dictionary ANTUSD, to the best of our knowledge. Like existing sentiment dictionaries, ANTUSD contains manually sentiment labels and scores from machine estimation. Moreover, ANTUSD provides stats from several manually labeling processes for further utilization, which can serve as reliable clues. We have conducted sentiment identification and classification experiments on more than 10 thousand words using this large dictionaries. With ANTUSD, we achieved the superior f-score 98.21% for polarity classification and 91.97% for opinion word extraction plus classification. Results show that with this large sentiment dictionary, simple classifiers can achieve good results, which is encouraging for further application development. In the future we will explore different approaches to link or inject more information to ANTUSD for it to serve as a better source of sentiment analysis features.

6. Acknowledgements

Research of this paper was partially supported by Ministry of Science and Technology, Taiwan, under the contract MOST104-2221-E-001-024-MY2

7. References

Collobert, R., & Weston, J. (2008). *A unified architecture for natural language processing: Deep neural networks with multitask learning.* Paper presented at the Proceedings of the 25th international conference on Machine learning.

Dong, Z., & Dong, Q. (2006). *HowNet and the Computation of Meaning*: World Scientific.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research, 9*, 1871-1874.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2014). Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*.

Harris, Z. S. (1954). Distributional structure. *Word, 10*(2-3), 146-162.

Huang, C.-R. (2009). Tagged Chinese Gigaword Version 2.0, LDC2009T14. *Linguistic Data Consortium*.

Huang, T.-H., Ku, L.-W., & Chen, H.-H. (2010). *Predicting Morphological Types of Chinese Bi-Character Words by Machine Learning Approaches.* Paper presented at the Proceedings of LREC.

Huang, T.-H. K., Chen, Y.-N., & Kong, L. (2015). ACBiMA: Advanced Chinese Bi-Character Word Morphological Analyzer. *ACL-IJCNLP 2015*, 26.

Ku, L.-W., Huang, T.-H., & Chen, H.-H. (2009). *Using morphological and syntactic structures for Chinese opinion analysis.* Paper presented at the Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3.

Ku, L.-W., Huang, T.-H., & Chen, H.-H. (2010). *Construction of a Chinese Opinion Treebank.* Paper presented at the LREC.

Ku, L.-W., Liang, Y.-T., & Chen, H.-H. (2006). *Opinion Extraction, Summarization and Tracking in News*

and Blog Corpora. Paper presented at the AAAI spring symposium: Computational approaches to analyzing weblogs.

Ku, L. W., Ho, H. W., & Chen, H. H. (2009). Opinion mining and relationship discovery using CopeOpi opinion analysis system. *Journal of the American Society for Information Science and Technology, 60*(7), 1486-1503.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality.* Paper presented at the Advances in neural information processing systems.

Mnih, A., & Hinton, G. E. (2009). *A scalable hierarchical distributed language model.* Paper presented at the Advances in neural information processing systems.