On Temporally Sensitive Word Embeddings for News Information Retrieval

Tae-Won Yoon
School of Computing
KAIST
Daejeon, South Korea
dbsus13@kaist.ac.kr

Sung-Hyon Myaeng
School of Computing
KAIST
Daejeon, South Korea
myaeng@kaist.ac.kr

Hyun-Wook Woo
Naver Corp.
Seongnam-si, South Korea
hw.woo@navercorp.com

Seung-Wook Lee
Naver Corp.
Seongnam-si, South Korea
swook.lee@navercorp.com

Sang-Bum Kim
Naver Corp.
Seongnam-si, South Korea
sangbum.kim@navercorp.com

Abstract

Word embedding is one of the hot issues in recent natural language processing (NLP) and information retrieval (IR) research because it has a potential to represent text at a semantic level. Current word embedding methods take advantage of term proximity relationships in a large corpus to generate a vector representation of a word in a semantic space. We argue that the semantic relationships among terms should change as time goes by, especially for news IR. With unusual and unprecedented events reported in news articles, for example, the word co-occurrence statistics in the time period covering the events would change non-trivially, affecting the semantic relationships of some words in the embedding space and hence news IR. With a hypothesis that news IR would benefit from changing word embeddings over time, this paper reports our initial investigation along the line. We constructed a news retrieval collection based on mobile search and conducted a retrieval experiment to compare the embeddings constructed from two sets of news articles covering two disjoint time spans. The collection is comprised of 500 most frequent queries and their clicked news articles in July, 2017, provided by Naver Corp. The experimental result shows there is a need for word embeddings to be built in a temporally sensitive way for news IR.

In: D. Albakour, D. Corney, J. Gonzalo, M. Martinez, B. Poblete, A. Vlachos (eds.): Proceedings of the NewsIR'18 Workshop at ECIR, Grenoble, France, 26-March-2018, published at http://ceur-ws.org

1 Introduction

The method of representing words and texts as vectors has drawn much attention in the natural language processing (NLP) and information retrieval (IR) areas. Various embedding methods for words, sentences, and paragraphs have emerged to represent them in a low dimensional vector space so that their semantic relationships can be computed[MSC+13, PSM14]. Miklov et al.[MSC+13] proposed two efficient word-level embedding models, Skip-gram and CBOW, both using an objective function to predict the relationship of words in a sentence. A different approach was proposed based on matrix factorization over a word-word matrix with a neural network model by Pennington et al.[PSM14]

One of the most important issues in building an embedding model is choosing an appropriate corpus for training. There have been several studies on the effect of employing different corpora for their types and domains in training embeddings. Siwei Lai at el.[LLHZ16] tested five different embedding models with three different domain corpora (wiki-dump, NYT corpus, IMDB corpus) on eight different tasks. They conclue that the influence of the domains is dominant in most tasks, proving the importance of choosing a

right domain. Diaz et al.[DMC16] also showed the importance of using a corpus with the same domain in a query expansion task by comparing different embedding spaces, one trained globally and the other trained on a local task-specific corpus. They used Skip-gram and Glove for embedding models, five different local corpora for retrieval and embedding training. They found that a locally trained embedding model works much better than globally trained one in the query expansion task.

Word embeddings may not reflect the dynamic nature of word meanings if a static collection is used for training. It is natural that new words coined with technological advances or emerging cultures can change the word embedding space. Especially in a news corpus that describes new events and contemporary issues, changes in word statistics would be more phenomenal and the word embedding space should also change accordingly. With an extensive coverage of an unusual real-life event in news articles, such as the terror in Las Vegas in 2017, the semantic distance between terms like Las Vegas and gun control, for example, would become much closer at least for a time being. We argue that capturing this type of word meaning dynamics should improve news IR and recommendation tasks.

While the aforementioned research showed the importance of considering the domain of the corpus, there has not been much work on investigating the importance of the publication time of the corpus for retrieval tasks. As time goes by, the meaning of a word and its relationship to other words would change, too. Kulkarni et al.[KARPS15] shows that as time goes by, the meaning and the usage of words changes. They analyze the change of word meanings and the relationship between words based on the time frames. However, they just focus on a computational approach to detect statistically significant linguistic shifts, and did not apply result to to retrieval tasks.

We examined the importance of the time periods of news corpora used for word embedding training by conducting a similarity-based news retrieval experiment based on three different corpora (Korean Wikipedia articles and news articles in March and in July, 2017) and two different commonly used word embedding models. A news retrieval collection was developed by extracting the most frequently asked 500 queries in July, 2017, and their clicked news articles in the click-through news data. For evaluation, we used the news retrieval task based on inverse document frequency weighted word centroid similarities (CentIDF), proposed by Georgios-Ioannis Brokos et al.[BMA16]. For each query in the retrieval experiment, we ranked the news documents based on the cosine similarity between the query embedding and a document embedding and compared the result against the gold standard constructed from the click-through data.

2 Models and Dataset

2.1 Embedding Models

We employed two most well-known word embedding models: word2vec (skip-gram version) proposed by Miklov et al.[MSC+13] and Glove by Pennington et al.[PSM14].

Word2vec. This model has two different versions, CBOW and Skip-gram, both of which use the context words of the target word to compute its semantics. CBOW uses the context words as the input and attempts to predict the target word from them. Skip-gram, on the other hand, calculates the probability of existence of the context words based on the target word. For optimization, a negative sampling method and hierarchical softmax function can be used. Negative sampling is an optimization method that uses not all the words but randomly sampled ones. Hierarchical softmax is a method that keeps all words mutual appearance information into a binary tree to reduce the calculation cost. In our work, we used Skip-gram with negative sampling[1].

Glove. This model is based on matrix factorization over a word-word matrix with a neural network model. It converts the word-word co-occurrence information to vectors. After training, the dot product of two words becomes proportional to the log value of concurrent probability of the two words. According to Pennington et al.[PSM14], the Glove model has been known to show a superior result in word analogy tasks and good at preserving semantic word relationships rather than syntactic ones.

2.2 Dataset

Click-through data. In order to evaluate the performance of multiple sets of word embeddings for the retrieval task, we employed a news corpus with news click-through data provided by Naver Corp.[2], the biggest portal service provider in South Korea, serving around 42 million users. The news click-through data covers all the mobile search clicks that took place between July 1 and July 9, 2017. The number of records or clicks is 53,472,390. The details of the test collection constructed from the click-through data is in section 3.2.1 below.

July news corpus. This corpus was generated from the news click-through data and used for training. All the clicked news articles were collected regardless of the number of clicks. When the embeddings were

[1]We also tested the CBOW model but the result is omitted because it shows similar tendency

[2]https://www.navercorp.com/en/index.nhn

constructed, only the nouns extracted from the news text were used. This corpus shares the same domain and the collection time with the retrieval evaluation collection. This corpus consists of 6,011,811 unique news articles with 1,232,910 tokens[3].

March news corpus. We collected news articles clicked in March, four months earlier than the period of the evaluation corpus, so that we can examine how the time difference affects the word embedding result in the news domain. Like the July corpus, only the nouns extracted by a morphological analyzer were used. This corpus has the same domain with the retrieval evaluation collection but a different time period. This corpus consists of 10,398,040 unique news articles with 1,381,901 tokens.

Wiki corpus. In order to reassure the importance of the training data domain, especially for news IR, we also built a collection of general articles from Korean Wikipedia and Namu-wiki, which are the most widely used online encyclopedic wiki collections in Korea. Like the news corpora, only the nouns were extracted and used for word embeddings. A Wikipedia dump (389,584 articles) and a Namu- wiki dump (533,406 articles) were downloaded in December 2017 and March 2017, respectively. Given that the test corpus was based on the queries in July, searching the Wikipeida documents generated at a later time until December gives the effect of searching future data (see Fig. 1). While this may seem irrational for news search, it should not affect the experimental result in that the Wikipeida articles are not so sensitive to time and that the number of future articles is relatively small. Namu-wiki played a more dominant role than Wikipedia in that the former contains more articles with a longer text per article. The total size of the Namu-wiki corpus is 4 times bigger than that of the wikipedia corpus. The resulting corpus contains 922,990 articles with 2,167,577 tokens in total.

Table 1: The dataset used for comparisons. All the data were collected in 2017.

Name	Domain	Collection Time	# Articles	# Tokens
Wiki corpus	Wiki	March,December	922,990	2,167,577
March news	News	March	10,398,040	1,381,901
July news	News	July	6,011,811	1,232,910

3 Experiment

The main goal of the experiment is to gain an insight on the need to use word embeddings computed from different time periods for news IR that usually seeks contemporary information, by comparing word embedding results from the three different types of corpora for a simple news retrieval task. As such, we do not attempt here to compare these embedding-based retrieval results against either word-based or embedding-based state-of-the-art IR methods. We make the retrieval process as simple as possible so that we can observe the effect of different embedding methods on the retrieval process without an interference of other factors that have been devised for retrieval effectiveness.

[3]All the datasets used in this paper are in Korean. They are used after extracting nouns based on the results from the morphological analyzer provided by Naver Corp. The examples of the terms given in this paper are English translations

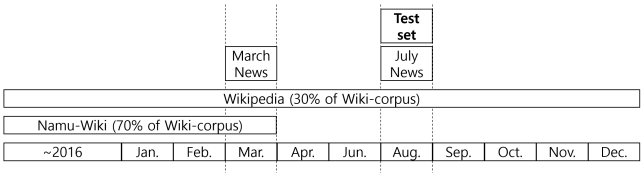

Figure 1: Date of each corpus used for the experiment. The time periods of the corpora used for the experiment. Even though one third of the Wikipedia documents were created after the test set, the future documents is only one tenth of the entire wiki corpus because the portion of Wikipedia corpus is only 30% of the whole and there four months after July.

3.1 Training and Parameter Settings

For the training of generating word embeddings, we used python gensim library[4] for word2vec and the author-provided code[5] for Glove. Other parameters for the Skip-gram model are: 300 for the vector dimension, 5 words for the context window size, and 0.0001 for the learning rate. For dropout, all words that appear less than 3 times were ignored. For Glove, we trained it with 300 for the vector dimension, 15 for the context window size, and 15 for maximum iterations. All words that appear less than 5 were dropped out.

3.2 Evaluation via News Retrieval

3.2.1 Evaluation-set

Based on the past research that claims using click-through data can be an alternative way to evaluate retrieval performance[J+03, LFZ+07], we selected 500 most frequently occurred queries from the news click-through data introduced in section 2.2. The queries were searched (or used) at least 6,000 times with the average of 36,521 times all the way up to about one million times. By taking a union of the clicked news articles, the resulting test collection consists of 500 queries and 17,530 documents that were clicked at least twice by the users who entered queries to the search engine. After excluding the news articles that

[4]https://radimrehurek.com/gensim/
[5]https://github.com/stanfordnlp/GloVe

were clicked just once, a query has 33.5 relevant documents on average with the maximum of 439.

3.2.2 Experimental Setup and Evaluation Metrics

To generate a vector for a query or a news article, we used the TF-IDF weighted word centroid calculation method (CentIDF[6]) proposed by Georgios-Ioannis Brokos et al.[BMA16]. A document vector $\overrightarrow{t}$ is computed as follows:

$$\overrightarrow{t} = \frac{\sum_{j=1}^{|V|} TF(w_j, t) \cdot IDF(w_j) \cdot \overrightarrow{w_j}}{\sum_{j=1}^{|V|} TF(w_j, t) \cdot IDF(w_j)}$$

Where $|V|$ is the vocabulary size of each sentence, w_j as a word at j-th position in the sentence t.

After generating document and query vectors, news articles are ranked according to cosine similarity with each query vector. The ranked list of news articles is used as a search result for the query. For comparisons among different embedding results, we use the result of three commonly used evaluation metrics: precision at 10, mean average precision (MAP) and NDCG at 10 based on binary relevance decisions.

3.2.3 Analysis of Retrieval Performance

The overall comparisons among the three different corpora are summarized in Table 2 for two different embedding models. For the Skip-gram model, the MAP result of the model trained on the July corpus is shown to perfrom 5.5% better than that trained on the March corpus although the time difference was only four months. The improvement was as high as 12% when compared to the result trained on a general corpus (the wiki corpus), i.e. on a different document type or domain. For the Glove model, the MAP result trained on the July corpus is shown to be about 5.5% better than both the model trained on the general corpus and the model trained on the March corpus. This strongly suggests that it is critical to build embeddings with the corpus in a similar time period for news retrieval.

The Skip-gram model is more sensitive to the domain than the Glove model. This is because the Glove model is better at extracting semantic relationships among words than syntactic ones. That is, the stylistic differences between the Wiki corpus and the March news corpus (without any temporal benefits) are less important. For the Skip-gram model, on the contrary, the writing style of the Namu-wiki corpus being sometimes informal with miscellaneous information and Internet slangs make the Wiki corpus result worse than the March corpus. This suggests that it is critical to build embeddings with the corpus in a similar domain and writing style when the Skip-gram model is used.

An important finding is that regardless of the metrics used, the July corpus gave the best results. While this is somewhat expected at an abstract level, it provides an important insight on the use of embeddings for IR. Using embeddings as opposed to words would increase recall, perhaps at the expense of lower precision in IR because of flexible matches. However, the experimental result shows increased precision with a more contemporary corpus used for embedding construction. This suggests that the embeddings constructed from the same time period better reflect the semantics of the words used by the users. Given that the embeddings capture the context of a target word, two words appearing in a close proximity in a corpus would share similar semantics. This would have the effect of retrieving news articles that may not have the exact query word (hence higher recall) and of reinforcing their relevance with the matched related words of the right context (hence high precision).

Table 2: Evaluating embedding models based on a news retrieval task. Bold faced numbers are the best results in different metrics. Both CentIDF and Arithmetic Mean are used for sentence embedding.

CentIDF			
Model	Precision@10	NDCG@10	MAP
Glove (wikipedia)	0.7114	0.7654	0.6192
Glove (March)	0.7046	0.7600	0.6188
Glove (July)	**0.7300**	**0.7776**	**0.6533**
Skip-gram (wikipedia)	0.6915	0.7509	0.5939
Skip-gram (March)	0.7203	0.7719	0.6317
Skip-gram (July)	**0.7399**	**0.7841**	**0.6666**
Arithmetic Mean			
Model	Precision@10	NDCG@10	MAP
Glove (wikipedia)	0.6015	0.6518	0.5138
Glove (March)	0.6023	0.6529	0.5263
Glove (July)	**0.6612**	**0.7018**	**0.5948**
Skip-gram (wikipedia)	0.5658	0.5193	0.4763
Skip-gram (March)	0.6706	0.7147	0.5866
Skip-gram (July)	**0.7090**	**0.7491**	**0.6404**

3.2.4 Qualitative Analysis

In order to better understand the effect of different corpora on embeddings and potentially on retrieval, we picked two time-sensitive queries corresponding to two separate sensational incidents in Korea between July 1 and July 9 and computed cosine similarity between the embedding of each and those of other words to rank them when the three different corpora were used. The first one was related to a claim made by several parents that McDonalds hamburgers caused a hamburger disease (Hemolytic uremic syndrome)[7], and the other

[6]It is known to be better than arithmetic mean. Unweighted method was also tried but without any gain.

[7]http://koreaherald.com/view.php?ud=20170705000868

Table 3: Top ten similar terms obtained by three different corpora for two sample queries "Hamburger disease (Hemolytic uremic syndrome)" and "Incheon kid murder". The expected intent-aware words are marked '*'

query: "Hamburger disease(Hemolytic uremic syndrome)"			query: "Incheon kid murder"		
Wiki corpus	March corpus	July corpus	Wiki corpus	March corpus	July corpus
Swing-top	215.8g	Hematotoxic*	Jung Duk Soon	Bupyeong	Murderer*
Substitute (food)	Burger*	Hemolytic*	Park Nari	Kidnap(while sleeping)	Elementary girl*
Cancer	Synchytrium endobioticum	Uremic*	Lee Duek Hwa	Before murder*	Final Verdict*
Soy-source bottle	Burger King	Basedow's disease	Woo Jung Sun	Doodle(river)	Killer*
Celiac spruse	Maclab	Shagas disease	Yang Jiseung	Taheutajeu	Don-Am dong
Basedow's disease	Beef	Maclab	Wentu Antu	After murder*	Park Chun Pung
Taste	Mayagbingsso	215.8g	Gak Jae Eun	Palda(mountain)	Incite Criminal*
Bread	Fast (food)*	Haemolyticity*	Oh Jong Guen	Siha(lake)	John Odgren
Parkinson's disease	BigKing	McDonald*	Song Yung Cil	Elementary girl*	Live-in lover
DOMDOM(burger)	Kim Kyo Bun	Uremicity*	Lee Wan Hue	Re-phase	Kidnap*

was the kidnap and murder of an eight-year-old girl in the elementary school by teenagers[8]. Table 3 shows top ten closest words under each corpus for the two queries.

For the "Hamburger disease" query, the result of Skip-gram trained on the wiki corpus consists of words that are generally related to each of the query words. Some are related to food (e.g. "Swing-top", "Substitute food", "Soy-source bottle", "Taste", "Bread") while others are to a disease (e.g. "Cancer", "Basedows disease", "Celiac pruse", "Parkisons disease"). But none of them are directly relevant to the intention of the query, such as "Hemolytic", "Umremic", and "McDonald". The result does not even contain words about "Burger" itself but those that are about the general notion of "Food" or "Disease". It is obvious that the embeddings constructed out of the Wiki corpus would bring in noise for news retrieval.

The result under the March corpus is completely different in the sense that the words about "Hambergur" were picked up. So the embedding space is much more focused on more contemporary issues in general. Since the Hamburger disease" related event didnt occur yet in March, however, none of the words are relevant to the query. It is very clear that the model trained on the July corpus gave the best result including the six intent-aware words with an asterisk.

For the "Incheon kid murder" query, the Skip-gram model trained on the wiki corpus gives a result consisting of perpetrators and victims of a murder in Korea, especially in Incheon, which would be good search terms if the intent were to retrieve general information, not about specific event-related news. It is because the target corpus contains articles about individual murder cases. On the other hand, the March corpus gave a completely different words that are related to descriptions of different murder cases, such as "Kidnap (while sleeping)", "before murder" and "After murder", contributing to the better retrieval result in the experiment. The model trained on the July corpus shows the most meaningful result containing six intent-aware words that would help retrieving relevant news articles.

While anecdotal, the examples in Table 3 constitute a strong indication that it is critical to use the corpus that coincides with the time-sensitive queries in news IR. The embedding space would be entirely different from that of the same news corpus covering a different time period, giving very different similarity relationships among words. As an example, we tested "presidential impeachment" as the query, which was a very sensational incident in March. We observe that the Skip-gram result trained on the wiki corpus has words that are unrelated to the query, such as president impeachment incident that took place in other countries, such as "Dilma Vana Rousseff", the former president in Brazil. The result under the March corpus is slightly better than the result under the July corpus since the incident took place at that specific time.

4 Conclusion and Future work

Given that timeliness is a rather unique aspect of new IR, word embeddings should be constructed in such a way that they reflect the evolving word-to-word relationships caused by emerging events and issues. Beginning with this hypothesis, we set out to build embeddings based on the news corpora of different time periods as well as on an encyclopedic corpus as a baseline for comparison, expecting to see the word embeddings constructed based on a temporally close corpus would help retrieving more relevant news articles than those based on temporally disparate documents.

We conducted an experiment with a newly constructed news IR corpus and a simple retrieval process using the cosine similarity measure for word embedding matches as well as qualitative analysis of the pseudo-expansion of query terms. The result clearly shows that it is worth constructing and using a corpus of temporally close news articles for news IR especially when word embeddings are used. The qualitative analysis of two sample queries strongly suggests that the semantic relationships among words change appropriately with different corpora so as to useful terms can be automatically generated for query expansion if the temporal and domain aspects of the corpora match

[8]http://koreaherald.com/view.php?ud=20170330000938

with the queries.

The initial result reported in the paper needs to be expanded in a number of different ways. Just to name a few, we first need to be able to suggest the appropriate time periods by which new embedding space must be created for news IR. Another immediate question is in what ways we can avoid new embedding constructions from the scratch when we have the embeddings for a series of past time spans. We are currently in the process of utilizing the past click-through data to capture the dynamic meaning changes across time periods.

Acknowledgment

This research was supported by the Naver Corp. and Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science & ICT (2017M3C4A7065963). Any opinions, findings and conclusions expressed in this material do not necessarily reflect the sponsors.

References

[BMA16] Georgios-Ioannis Brokos, Prodromos Malakasiotis, and Ion Androutsopoulos. Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing, BioNLP@ACL 2016, Berlin, Germany, August 12, 2016*, pages 114–118, 2016.

[DMC16] Fernando Diaz, Bhaskar Mitra, and Nick Craswell. Query expansion with locally-trained word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics(ACL), August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

[J+03] Thorsten Joachims et al. Evaluating retrieval performance using clickthrough data., 2003.

[KARPS15] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web(WWW)*, pages 625–635. International World Wide Web Conferences Steering Committee, 2015.

[LFZ+07] Yiqun Liu, Yupeng Fu, Min Zhang, Shaoping Ma, and Liyun Ru. Automatic search engine performance evaluation with click-through data analysis. In *Proceedings of the 16th international conference on World Wide Web*, pages 1133–1134. ACM, 2007.

[LLHZ16] Siwei Lai, Kang Liu, Shizhu He, and Jun Zhao. How to generate a good word embedding. *IEEE Intelligent Systems*, 31(6):5–14, 2016.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.