Learning Query Paraphrasing from A Monolingual Corpus

Meng ZHAO†, Hiroaki OHSHIMA†, and Katsumi TANAKA†

† Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshida Honmachi, Kyoto, 606–8501 Japan
E-mail: †{zhao,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract We propose a method to obtain different expressions that convey the same meaning with that of a given phrase or sentence query. The notion is based on the distributional hypothesis that words that occur in the same contexts tend to have similar meanings. We assume that a phrase or sentence can be divided into a template part and an entity part. Therefore, surrounding entities generate the context of a template and vice versa, surrounding (in precise, suitable) templates generate the context of an entity. In this paper, we present an extension to the continuous bag-of-words (CBOW) model that efficiently learns the embedding for each template. As a result, semantically similar templates are detected in such a way that their semantic representations are similar to each other.

Key words query paraphrasing, CBOW model, semantic representation

1. Introduction

Paraphrases are linguistic expressions that restate the same meaning using different variations. In the most extreme case, they may not be even similar in wording. It has been shown that paraphrases are useful in many applications. For example, paraphrases can help detect fragments of text that convey the same meaning across documents and this can improve the precision of multi-document summarization [6] [16]. In the field of machine translation, [10] [14] [15] show that augmenting the training data with paraphrases generated by pivoting through other languages can alleviate the vocabulary coverage problem. In information extraction, [33] [11] [9] present approaches incorporating paraphrases to extract semantic relations among entities. In information retrieval, paraphrases have been used for query expansion [26] [1] [32]. A large proportion of previous work extract and generate paraphrases based on parallel corpora [5] [2] or comparable corpora [4] [27] [30]. However, there are limitations in using those corpora. For example, the quality of obtained paraphrases strongly depends on the quality of the corpus, a high-quality corpus can cost a great deal of manpower and time to construct. Moreover, it may be hard to cover all possible genres. For example, [30] uses a corpus consisted of newswire articles written by six different news agencies.

A bottleneck for information retrieval systems is the identification of different expressions with the same meaning between a user's query and existing descriptions in the systems. In this paper, we aim at paraphrasing user-given queries to close to the expressions used in the systems, and consequently improving search performance. We introduce an unsupervised learning approach which is based on the word embedding technique, to capture paraphrases on a monolingual corpus.

Word embedding, where words or phrases from the vocabulary are mapped to dense vectors of real numbers, has attracted a lot of research interest, since it is first introduced by Bengio et al. [7] in 2003. It has been exceptionally successful in many natural language processing tasks such as named entity extraction [21] [25] [24], dependency parsing [3], machine translation [34] [18] and learning representations of paragraphs and documents [13].

With the help of word embedding, words whose meaning is similar are mapped to a similar position in the vector space. For example, word "strong" is similar to "powerful", word "Paris" is similar to "Berlin" [20]. Moreover, the learned distributed vectors can capture many linguistic regularities and patterns. For example, vector operations vector("Paris") - vector("France") + vector("Germany") results in a vector that is very close to vector("Berlin"), and vector("brother") - vector("grandson") + vector("granddaughter") is close to vector("sister") [17] [20].

However, the meaning of a phrase cannot be obtained by simply combining the meanings of words in the phrase. For example, the meaning of "Volga River" cannot be obtained by combining the meanings "Volga" and "River". Mikolov el al. realized the problem and hence proposed a data-driven approach where phrases are formed based on the unigram and bigram counts to find phrases and then phrases are re-

garded as individual tokens for the later training [19].

We assume that a phrase or sentence can be divided into a template part and an entity part. Take a sentence "Bayer said it would buy Monsanto" for example. The template part corresponds to "said it would buy", while the entity part corresponds to a tuple (Bayer, Monsanto). Alternative expressions for either the template part or the entity part can be found, and hence a variety of combinations could be generated. Here, we concentrate on learning alternative expressions for templates, where the length for both of them are variable. For example, vector("has announced its acquisition of") is close to vector("said it would buy"). We find that still, it is difficult to capture the meanings of variable-length pieces of texts by the above approach [19]. Therefore, we present a method based on the continuous bag-of-words (CBOW) model [17] which predicts multiple words rather than a single word based on the context. The learned vectors of words in a template are summed, averaged or concatenated to represent the meaning of the template.

The remainder of the paper is organized as follows. In Section 2., we briefly review the CBOW model. In Section 3., we describe the details of our extension to the CBOW model. In Section 4., we show experimental results, as well as some interesting observations. We dedicate Section 5. to the discussion of the previous work on paraphrase detection and acquisition, adopting machine learning approaches. Finally, in Section 6., we conclude the paper and draw some future directions.

2. Background: Continuous Bag-of-words Model

The CBOW model [17] learns distributed vector representation for words by predicting a word given the surrounding words as the context. The model architecture is shown in Figure 1. x_i is a one-hot vector with dimensionality V and consists of 0s in all cells with the exception of a single 1 in the cell corresponding to the i-th word, denoted by t_i. Here, V is the size of vocabulary. W is the input to hidden weight matrix, W' the hidden to output weight matrix.

In the CBOW model, $v(t) \in R^N$ is the vector representation of the word $t \in T$, where T is the word vocabulary and N is the number of hidden units, thus the dimensionality of the word embeddings. $v(t_i)$ is easily obtained by left multiply x_i by the input to hidden weight matrix W. Given a training set including the sequence of words t_1, t_2, t_3, ..., t_V, the word embeddings are learned by maximizing the average log probability

$$\frac{1}{V}\sum_{i=c}^{V-c} p(t_i|t_{i-c}, ..., t_{i+c}) \qquad (1)$$

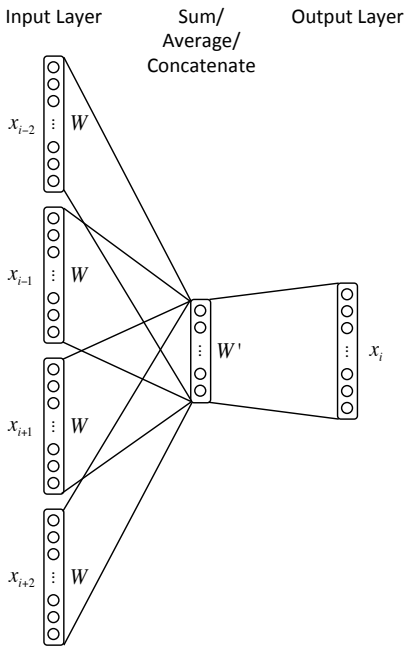

Figure 1 Architecture of the CBOW model with window size $c = 2$. x_i is a one-hot vector with dimensionality V and consists of 0s in all cells with the exception of a single 1 in the cell corresponding to the i-th word, denoted by t_i. Here, V is the size of vocabulary. Context of word t_i consists of words t_{i-2}, t_{i-1}, t_{i+1} and t_{i+2}.

where c is the window size of the context.

Basically, $p(t_i|t_{i-c}, ..., t_{i+c})$ is defined by using the softmax function

$$p(t_i|t_{i-c}, ..., t_{i+c}) = \frac{exp(\hat{v}_{t_i}^\top \cdot v_{c_{t_i}})}{\sum_{t \in T} exp(\hat{v}_t^\top \cdot v_{c_{t_i}})} \qquad (2)$$

where v_t and $\hat{v}_t$ are the input and output vector representations of t, and $v_{c_{t_i}}$ denotes the sum, average or concatenation of context words.

Because the cost of computing the log probabilities of all words is very expensive, hierarchical softmax [22], an approximation inspired by binary trees, is employed for fast training. Mikolov et al. used a binary Huffman tree for speedup in a series of their research [19] [18] [13].

3. An Extension to the CBOW Model

To extend the CBOW model to learn embeddings of templates, we use the surrounding words to predict multiple words rather than a single word. The model architecture is shown in Figure 2 where for simplicity, two words are predicted given the context. t_{i_f} and t_{i_l} indicate the former and latter words, respectively, considering the relative position in a training sequence. We denote M the maximum distance between t_{i_f} and t_{i_l}. If the distance between them is larger than M, we discard the probability that they co-occur. Figure 3 shows an example of context words in the model, where

"Google" denotes t_{i_f} and "DoubleClick" denotes t_{i_l}.

Given a training set including the sequence of words t_1, t_2, t_3, ..., t_V, the word embeddings are learned by maximizing the average log probability

$$\frac{1}{V(V-1)} \sum_{t_{i_f}, t_{i_l} \in T-C, t_{i_f} \neq t_{i_l}} p(t_{i_f}, t_{i_l} | t_{i_f-c}, ..., t_{i_l+c}) \quad (3)$$

where t_{i_f} and t_{i_l} are not identical, C is the set of context words and c is the window size of the context. We use t_O to represent $\{t_{i_f}, t_{i_l}\}$, t_I to represent $\{t_{i_f-c}, ..., t_{i_l+c}\}$ for simple notation hereinafter. As a result, $p(t_{i_f}, t_{i_l} | t_{i_f-c}, ..., t_{i_l+c})$ is equal to $p(t_O|t_I)$.

We define $p(t_O|t_I)$ using the softmax function

$$p(t_O|t_I) = \frac{exp(\hat{v}_{t_O}^{\top} \cdot v_{t_I})}{\sum_{t \in T \times T} exp(\hat{v}_t^{\top} \cdot v_{t_I})} \quad (4)$$

where v_t and $\hat{v}_t$ are the input and output vector representations of t, respectively.

We define the distributed vector representation of a template as the average of template word vectors weighted by the input to hidden weight matrix W:

$$\frac{1}{L} \sum_{t_i \in template} W \cdot x_i \quad (5)$$

where L denotes the number of words contained in a template. Therefore, the vector representation of "plan to acquire" is the average of vectors of the words "plan", "to" and "acquire".

4. Experiments

We plan to train the model on the full English Wikipedia text(注1).

To evaluate whether templates with the same, or similar meaning could be mapped to a similar position in the vector space, we plan to select some relations by referring to some previous works [8] [9] [12] about acquiring paraphrases or detecting paraphrases in a corpus, and questions from TREC-8 Question-Answering Track.

5. Related Work

Paraphrase acquisition is a task of acquiring paraphrases of a given text fragment. Some approaches have been proposed for acquiring paraphrases at word, or phrasal level. However, these techniques are designed to be only suitable for specific types of resources. Shinyama et al. [28] and Wubben et al. [31] acquired paraphrases from news articles. For example, Shinyama et al. [28] argued that news articles by different news agents reporting the same event of the same day

(注1)：It can be downloaded from `http://en.wikipedia.org/wiki/Wikipedia_database`.

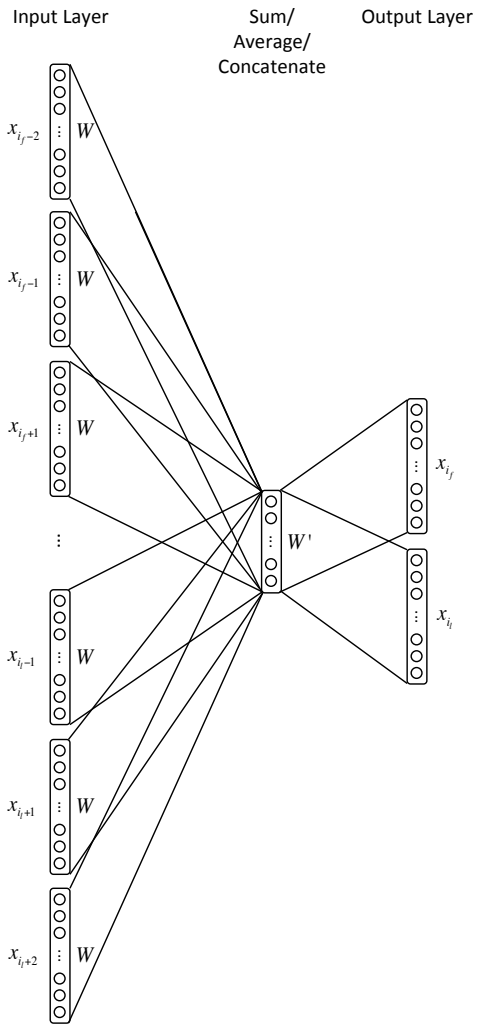

Figure 2 Architecture of the extended CBOW model with window size $c = 2$. Context of words t_{i_f} and t_{i_l} consists of three parts: words before t_{i_f} with window size c, i.e., t_{i_f-2}, t_{i_f-1}, words between t_{i_f} and t_{i_l}, words after t_{i_l} with window size c, i.e., t_{i_l+1} and t_{i_l+2}.

can contain paraphrases. Thus, they proposed an automatic paraphrase acquisition approach based on the assumption that named entities are preserved across paraphrases.

Paşca and Dienes proposed a different method [23]. They use inherently noisy, unreliable Web documents rather than clean, formatted documents. They assumed that if two sentence fragments have common word sequences at both extremities, then the variable word sequences in the middle are potential paraphrases of each other. Socher et al. proposed an unsupervised feature learning algorithm based on recursive autoencoders(RAE) in [29] for paraphrase detection. Their RAEs are based on a novel unfolding objective and learn feature vectors for phrases in syntactic trees. These features are used to measure the word- and phrase-wise similarity between two sentences.

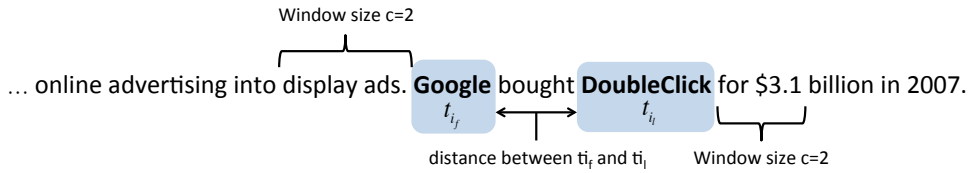

Figure 3 An example of context words for t_{i_f} and t_{i_l} with window size $c = 2$.

6. Conclusion

We proposed a method which is an extension to the CBOW model, to obtain different expressions that convey the same meaning with that of a given phrase or sentence query. We assume that a phrase or sentence can be divided into a template part and an entity part. Especially, we focused on the template part. After the training, templates with the same, or similar meaning are mapped to a similar position in the vector space.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers JP24680008, JP15H01718, JP16H02906.

References

[1] Anick, P.G., Tipirneni, S.: The paraphrase search assistant: Terminological feedback for iterative information seeking. In: Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 153–159 (1999)

[2] Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 597–604 (2005)

[3] Bansal, M., Gimpel, K., Livescu, K.: Tailoring continuous word representations for dependency parsing. In: ACL (2). pp. 809–815 (2014)

[4] Barzilay, R., Elhadad, N.: Sentence alignment for monolingual comparable corpora. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. pp. 25–32 (2003)

[5] Barzilay, R., McKeown, K.R.: Extracting paraphrases from a parallel corpus. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. pp. 50–57 (2001)

[6] Barzilay, R., McKeown, K.R., Elhadad, M.: Information fusion in the context of multi-document summarization. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 550–557 (1999)

[7] Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. journal of machine learning research 3, 1137–1155 (2003)

[8] Bhagat, R., Ravichandran, D.: Large scale acquisition of paraphrases for learning surface patterns. In: Proceedings of ACL2008: HLT. pp. 674–682 (2008)

[9] Bollegala, D.T., Matsuo, Y., Ishizuka, M.: Relational duality: Unsupervised extraction of semantic relations between entities on the web. In: Proceedings of WWW. pp. 151–160 (2010)

[10] Callison-Burch, C., Koehn, P., Osborne, M.: Improved statistical machine translation using paraphrases. In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. pp. 17–24 (2006)

[11] Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Commun. ACM 51(12), 68–74 (Dec 2008)

[12] Idan, I.S., Tanev, H., Dagan, I.: Scaling web-based acquisition of entailment relations. In: Proceedings of EMNLP. pp. 41–48 (2004)

[13] Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: ICML. vol. 14, pp. 1188–1196 (2014)

[14] Madnani, N., Ayan, N.F., Resnik, P., Dorr, B.J.: Using paraphrases for parameter tuning in statistical machine translation. In: Proceedings of the ACL Workshop on Statistical Machine Translation (2007)

[15] Marton, Y., Callison-Burch, C., Resnik, P.: Improved statistical machine translation using monolingually-derived paraphrases. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. pp. 381–390 (2009)

[16] McKeown, K.R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J.L., Nenkova, A., Sable, C., Schiffman, B., Sigelman, S., Summarization, M.: Tracking and summarizing news on a daily basis with columbia's newsblaster. In: Proceedings of the second international conference on Human Language Technology Research. pp. 280–285 (2002)

[17] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

[18] Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. CoRR abs/1309.4168 (2013)

[19] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems. pp. 3111–3119 (2013)

[20] Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: HLT-NAACL. vol. 13, pp. 746–751 (2013)

[21] Miller, S., Guinness, J., Zamanian, A.: Name tagging with word clusters and discriminative training. In: North American Chapter of the Association for Computational Linguistics: Hu- man Language Technologies (NAACL-HLT). vol. 4, pp. 337–342 (2004)

[22] Morin, F., Bengio, Y.: Hierarchical probabilistic neural network language model. In: Proceedings of the International Workshop on Artificial Intelligence and Statistics. pp. 246–252 (2005)

[23] Paşca, M., Dienes, P.: Aligning needles in a haystack: Paraphrase acquisition across the web. In: Proceedings of IJC-

NLP. pp. 119–130 (2005)
[24] Passos, A., Kumar, V., McCallum, A.: Lexicon infused phrase embeddings for named entity resolution. arXiv preprint arXiv:1404.5367 (2014)
[25] Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning. pp. 147–155 (2009)
[26] Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc. (1986)
[27] Shinyama, Y., Sekine, S.: Paraphrase acquisition for information extraction. In: Proceedings of the Second International Workshop on Paraphrasing. vol. 16, pp. 65–71 (2003)
[28] Shinyama, Y., Sekine, S., Sudo, K.: Automatic paraphrase acquisition from news articles. In: Proceedings of HLT. pp. 313–318 (2002)
[29] Socher, R., Huang, E.H., Pennington, J., Ng, A.Y., Manning, C.D.: Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In: NIPS. vol. 24, pp. 801–809 (2011)
[30] Wang, R., Callison-Burch, C.: Paraphrase fragment extraction from monolingual comparable corpora. In: Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web. pp. 52–60 (2011)
[31] Wubben, S., van den Bosch, A., Krahmer, E., Marsi, E.: Clustering and matching headlines for automatic paraphrase acquisition. In: Proceedings of ENLG. pp. 122–125 (2009)
[32] Yamamoto, Y., Tanaka, K.: Towards web search by sentence queries: Asking the web for query substitutions. In: Proceedings of the 16th International Conference on Database Systems for Advanced Application (DASFAA 2011). pp. 83–92 (2011)
[33] Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., Soderland, S.: Textrunner: Open information extraction on the web. In: Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. pp. 25–26 (2007)
[34] Zou, W.Y., Socher, R., Cer, D.M., Manning, C.D.: Bilingual word embeddings for phrase-based machine translation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1393–1398 (2013)