Word Embedding for Cross-lingual Natural Language Analysis

Yukun Hu *

University of California, Irvine California, USA

* Corresponding Author Email: Yukunh2@uci.edu

Abstract. Word embedding, a distributed representation of natural language based on deep neural networks, has made significant breakthroughs in many natural language processing tasks and has gradually become a hot subject in research and application. Word embedding methods can capture more complex and valuable semantic information than existing methods. However, existing methods of word embedding often rely on large-scale annotation resources, which are often difficult to obtain, especially for resource-poor languages. In response to this problem, researchers have explored different research routes, such as unsupervised learning from untagged data, semi-supervised learning that integrates tagged and untagged data, or crowdsourcing. At the same time, many scholars have proposed to improve the analysis accuracy of target tasks by integrating the annotation resources of different languages and enabling knowledge from foreign languages to be transferred or merged with models. This paper discusses the development and prospects of word embedding.

Keywords: Word embedding, Nature Language Processing, Deep learning, Machine Learning.

1. Introduction

Representing language into forms that computers can process is the premise and key to natural language processing, and preserving as much contextual information as possible can make the results of natural language processing more accurate [1-3]. Existing natural language representation methods can be divided into two types: 1) Discrete forms of symbolic representation, such as the one-hot representation of words; 2) Distributed representation; Wherein the one-hot representation of words represents each word as a very long vector, and the dimension of the vector is the size of the word list. And only the value of the dimension of the current word is 1, and the importance of the remaining dimensions is 0, which is simple and intuitive. However, the dimension of the vector increases as the word list increases, and this representation method makes any two words isolated, which cannot express the relevant information between words at the semantic level. To incorporate context into word representation, researchers proposed a distributed representation of multiple words based on the distributed hypothesis[4]. In addition, the word embedding method has been proposed to solve the problem of excessive dimensions in recent years, based on the distribution representation of neural networks [5-7]. Compared with the one-hot representation, the word embedding method can preserve more contextual information while reducing the dimensionality of the data. The most representative method in this category is word2vec [6]. At present, the word embedding method has been able to preserve semantic information well and reduce the dimension of the data.

However, there are more than 5,000 languages in the world, of which there are about 150 languages with more than 1 million speakers. However, for most natural language processing tasks, the scale of annotation resources in different languages shows an extremely serious long tail phenomenon. Taking syntactic analysis, which is difficult to tag, as an example, there are only more than 50 languages with a certain size of syntactic treebanks. Therefore, the study of cross-lingual knowledge transfer methods for resource-poor languages has gradually become an important research subject. Currently, better-performing methods of Cross-lingual word embedding still rely heavily on bilingual parallel resources, and large-scale bilingual data is challenging to obtain for most truly resource-poor languages [8]. Fortunately, there is still a large amount of data of other data types available to researchers. Although these data and the text type of corpora are pretty different in the data type (such as pictures), the domain knowledge in these data has strong commonalities and may complement the application. Besides, bias is becoming an increasingly severe problem in natural language processing, and even word

embeddings trained on Google News articles show this trend, such as inherent biases about gender [9]. In the process of cross-lingual knowledge transfer, Bias also transfers with it. Understanding what biases are acquired by word embeddings and finding good ways to remove them is critical for developing fair and unbiased natural language processing applications.

This paper reviews the knowledge transfer method based on cross-lingual word embedding, which aims to build a bridge between different languages for information exchange and knowledge transfer so that we can make full use of data from different languages and different structures to improve the analysis accuracy of the target language or specific tasks.

2. Methods

2.1. Vectorized representation of natural language

In natural language processing based on statistical models, the common idea is to first extract features for the target object so that each case is represented as a feature vector and used as input to the statistical model. Then, statistical machine learning methods estimate the parameters in the model. Feature representation is a fundamental part of natural language processing research and one of the critical factors affecting the performance of machine learning systems. The available vectorized representations of natural language text are classified into two types.

Discrete forms of symbolic representations, such as One-Hot Representation of words and Bag-of-Words representations of documents. Take the One-Hot Representation of words as an example. It represents each word as a very long vector. The dimension of the vector is the size of the word list, only the value of the dimension of the current word is 1, and the values of the remaining dimensions are 0. This representation method is intuitive and concise, and easy to calculate. Still, its disadvantage is that it only isolates the representation feature itself without portraying the semantic information it contains, so it cannot fully express the semantic associations between different symbolic data. In addition, the feature vectors obtained in this way tend to be of high dimension, limiting the complexity of the model.

Distributed representation. It is usually a continuous, dense low-dimensional vector representation. These methods are based on the semantic distribution assumption proposed by Firth- the meaning of a word is determined by the context in which it co-occurs, common natural language distributed representation methods include Latent Semantic Analysis (LSA), Max-margin hinge loss (MMHL), Skip-gram with negative sampling (SGNS), Continuous bag-of-words (CBOW) [10,11]. This kind of method can be subdivided into the following three categories.

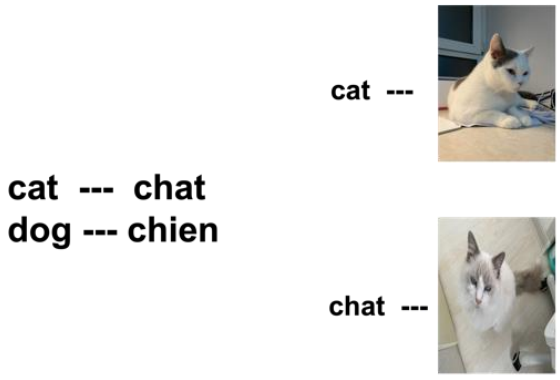
- First, matrix-based distribution representation; The matrix-based distribution representation is often referred to as the semantic distribution model. This method is usually by building a co-occurrence matrix (each row of the matrix corresponds to a word, and each column represents a context). The value of each element is the number of co-occurrences of the corresponding word and context in the corpora. Therefore, each word can be represented by the corresponding row vector in the matrix, and the similarity of their vectors can directly measure the similarity of any two words. Representations based on co-occurrence matrices are generally of high dimension and very sparse, making them difficult to apply directly to various natural language processing tasks. The use of dimensionality reduction techniques allows them to be converted to relatively low-dimensional, dense vectors and reduce the impact of noise.
- The second is cluster-based word representation; another commonly used distributed representation can be generated from clustering results, and its representative work is the algorithm proposed by Brown in 1992, called Brown Clustering [12]. This algorithm is a hierarchical clustering method in which words are represented by multi-layered categories resulting from clustering. The semantic similarity of any two words can be judged based on their common categories. The central idea of brown clustering is that the category of the current word is affected by the category of its antecedents, and semantically similar words have antecedents with similar categories. The Brown clustering only considers the semantic effects of the antecedent word, i.e., it uses only the antecedent

word as contextual information. Cluster-based distribution representations can also be changed from the co-occurrence matrix of words. Pereira et al. proposed a method to convert the co-occurrence matrix into a clustering distribution and thus represent the semantics.

- The third is neural network-based word representation; neural network-based word representation is generally called word vector or word embedding. This method maps each word to a low-dimensional, dense real vector. Each dimension of the vector represents the potential characteristics of the word, and the feature can capture proper grammar and semantic properties. The most significant advantage of word vectors is that we can judge the similarity of words relatively quickly and efficiently based on the distance between vectors (e.g., cosine distance) because semantically or grammatically similar words have similar vector representations. The construction of word vectors relies mainly on neural network models. Neural networks can represent arbitrary n-gram word groups in a flexible and linear combinatorial manner. The parameters only grow at a linear rate. The word vectors can capture more complex and valuable semantic information than the matrix-based distribution representation. The neural network-based word representation is the most advanced and effective vectorized representation of natural language.

2.2. Cross-lingual word embedding methods

The representation of different languages may vary greatly, but the representation of this information is uniform at the semantic level. For example, some researchers have noticed that the geometric relationships between words are similar between different languages. Based on the above views, researchers have proposed a variety of cross-lingual word embedding methods in recent years. According to the different target objects, the existing cross-lingual word embedding technologies can be divided into word-level alignment models, sentence alignment models, and document alignment models.



(a) Word, par. (b) Word, comp.

Figure 1. Examples for the nature and type of alignment of data sources. Par.: parallel. Comp.: comparable.

Word-level alignment models. Word-level alignment models is a cross-lingual word embedding model at the word level, usually using a one-to-one corresponding bilingual corpora (Figure 1 (a)) or other data that can semantically correspond to words one-to-one, such as pictures, etc. (Figure 1 (b)), through training, obtain the mapping relationship from the source object to target object semantically, as classified below.

The word-level alignment models based on bilingual parallel resources. The word-level alignment models based on bilingual parallel resource first needs to obtain the corpora resources of different languages, then obtain the word vectors of the single language separately, and finally obtain the mapping relationship from the source language to the target language through training. The specific steps are as follows, with the source language S^1,…, S^n , and the word vector of a one-to-one corresponding target language T^1,…, T^n as the training set, and then use the random gradient descent to convert the learning problem of the linear mapping matrix W into an optimization problem, and obtain the optimal solution by solving the MSE (mean squared error), as shown in Equation (1).

$$\Omega_{MSE} = \sum_{i=1}^{n} ||WS^i - T^i||_F^2 \quad (1)$$

Further, the basic word-level alignment models based on bilingual corpora can be formalized into Equation (2).

$$J = L^S + L^T + \Omega_{MSE} \quad (2)$$

Where $L^S + L^T$ is the loss function of the single-language word embedding model of the source language *S* and the target language T. For the above model, the researchers have made many extensions, which are generally aimed at building a better knowledge base of bilingual corpora or finding better mapping methods, and the basic principle has not changed.

Semantic-based heterogeneous data word-level alignment method. The existing semantic-based heterogeneous data word-level alignment methods are mainly studied for two types of data:

First, the picture data containing semantic information, as shown in Figure 1 (b); the basic principle of this type of method is to use the semantic space of the picture as a bridge to connect different languages, because the image description of the same thing in different languages is semantically the same, such as using the Google Picture search engine, whether you enter 'bicycle', 'fiets', 'Fahrrad' or other languages representing the bicycle, you can get a bunch of pictures containing bicycles.

Second, POS tage equivalence; the starting point of this type of method assumes that the text of different languages is roughly the same in the order of part-of-speech combinations, and then uses the part-of-speech tag corpora or its context of the source and target languages as a bridge to learning the semantic mapping relationship of different languages.

Sentence-level alignment models. Despite being in the era of big data, sentence-level bilingual parallel corpora data that can be used is still not easily available, and most of the existing studies are based on many sentence-level bilingual corpora data provided by Machine Translation. Like the word-level alignment models, the existing sentence alignment models can also be divided into two categories, of which the sentence alignment model based on bilingual corpora can be divided into four categories, as shown in Figure 2.

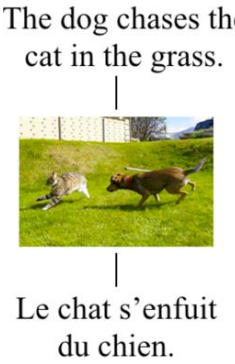


Figure 2. the overview of sentence alignment [8]

The sentence alignment model based on bilingual parallel resources. Word-level alignment method based on matrix decomposition. This kind of method first obtains the word embedding matrix X^S and X^T of the source language S and the target language T by training, and then obtains the mapping relationship between the source language and the target language sentence corpora by establishing two mapping matrices in different directions (i.e., sentence alignment is decomposed into word-level alignment or word- sentence alignment). The calculation formula is as follows:

$$\Omega_{S \to T} = ||X^T - A^{S \to T} X^S||^2 \quad (3)$$

$$\Omega_{T \to S} = ||X^S - A^{T \to S} X^T||^2 \quad (4)$$

Like the word-level alignment methods based on the bilingual corpora, the basic model of such methods can be formalized as follows:

$$J = L^S + L^T + \Omega_{S\to T} + \Omega_{T\to S} \tag{5}$$

Where, $L^S + L^T$ is the loss function of the single language word embedding model of the source language S and the target language *T*.

Combined sentence model. This type of methods first sums the word vectors of each word that make up the sentence (source language $sent^s$ and target language $sent^t$) to obtain the sentence vector representing the sentence. Then, through the optimization problem, using the way of that, the closer the distance between similar sentences is, the farther the distance between different sentences is, to obtain the optimal mapping method.

Bilingual automatic coding and decoding models. Different from the above two methods, the Bilingual automatic coding and decoding models no longer solves the minimum distance between the source language and the target language but reconstructs the sentence itself and its translation by coding and decoding the sentence corpora.

Bilingual skip-gram model. This type of methods is based on the bilingual bag-of-words model. Instead of minimizing the distance between two-sentence corpora pairs, this method minimizes the distance between word vectors in a two-sentence corpora pair.

Semantic-based heterogeneous data sentence alignment models. Like the semantic-based heterogeneous data word-level alignment models, the semantic-based heterogeneous data sentence alignment models are primarily aimed at image data that contains semantics. Existing sentence alignment models based on similar data usually use deep learning models, to extract semantic descriptions of images, and then effectively combine the knowledge from image and text data.

Document-level alignment models. Existing document alignment models can be divided into three categories: First, the Pseudo-bilingual document-aligned corpora. This type of methods typically assumes that documents in two different languages are structurally similar, and then insert words between the source and target documents according to the order of word occurrences in the monolingual documents and the length ratio of the monolingual documents. Since the majority of word embedding methods are context-based embedding learning, by establishing a bilingual corpus, each word has a robust bilingual context, thereby establishing an association between different languages.

Concept-based models based on the assumption that different languages use roughly similar texts to describe the same conceptual subject. In these methods, It usually first constructs the corpora of a document with a subject into a conceptual matrix, and learning a shared bilingual semantic space through a bilingual probability model. Extensions of sentence-alignment models learns the paragraph vector of the entire paragraph, and then, based on a method similar to the sentence alignment models, evaluates, rotates, and scales the semantic space of the source paragraph until it is closest to the semantics of the target paragraph. In short, it is to treat the entire paragraph as a sentence.

3. Discussion

In this part of the language study, we are talking about the language that has a significant role in common. The more common have between the two languages, the easier and more accurate our study and transfer can be from large systems to particular words. We can analyze all kinds of languages by language family, sub-family, branch, and sub-branch. If two languages are in the same language family but not the same branch, we will call it a cousin language; if they are in the same branch, we call them sister language. Based on our more common, more accurate rule. We can always say that learning and transferring between sister languages can be more precise than learning cousin languages. There are seven primary language families in our commonly known— the Indo-European family, the Uralic family, the Dravidian family, the Caucasian family, the Altaic family, the Sino-Tibetan, and the Austro-Asiatic family. When we are dealing with languages in the same family, the most direct way

is to do a word-to-word transfer and learning. The only material matters here is a large enough one-to-one data set. And the machine will have helped us do most of the work and left a super small percent of work that needs to be handled by the real person. Super evident that the benefit of this way is that the work can be directly done, and the map relation is one-to-one without further coding except for some small hand-made put. However, the negative affective is also apparent here: polysemy use in sentences and lack of accuracy if the relationship gap between languages is cousin languages or more. Facing these problems, we might need to set up not only a one-to-one code; but multiple characters maps that add together to make the accuracy of the work better.

In the newest update, we constantly face difficulties due to the subject. At the same time, the relationship between languages is too big or when a word has several meanings, and there might be colossal diffraction when we collect data from different backgrounds. If the two languages came from other language families: for example, Chinese and English. Chinese is from the Sino-Tibetan family; most of the words and phrases come from hand drawing pictures since all the languages from the family are pictographs. The characters were made up from how the things look or how the feelings make things look. But for English from the Germanic branch under the Indo-European family. The words are based on thousands of millions of etyma and affixes. In this situation, on the one hand, we have a language system based on an outlook of surroundings; on the other hand, we have a language system base on characters permutation and the logic relation when affixed with different sub meanings band more uncompromising. There seems complicated to make a machine understand what is happening when we want him to learn the relationship between them. However, when we take a step back, no matter what kind of language we are using and where all those words came from, the goal of using language should always be to help people communicate and let others understand them. Here is the benefit of why we are combining machine word learning with pictures. Because there will always be a word or sentence to describe the seeing no matter what kind of language. The engine will combine the characters into a matrix of the image. With this path, we can translate any language into another by pictures after making the machine learn all these words in combined views. Even further, when we can take a step forward—compare the picture to a sentence of description, we can let the machine learn more than just a word but also the relation between words that come up in a sentence.

The most well-known problem in the area nowadays is cultural bias—it commonly happens when the machine tries to learn from a non-native source. It doesn't mean that the reference is not kind or can't be used, but we should set up a self-checking code and a database to make the machine ignore to repeat studying the partial data and be a fair translator. For example, when you ask an American about the name of Chinese dishes, they will generally come up with orange chicken, but many Chinese people might never hear it before. If we ignore these parts in our work, the result of the data will always put words and expressions near the improper axis, causing dozens of misunderstandings while doing the transfer. Also, the context can substantially differ under different cultural backgrounds even though the exact meaning sentence is put in the same place. It will be super valuable if we can figure out a way to make our machine pre-learned some language environment background or set up a way to let the computer work out if the writer is sarcastic or just in a severe tone. Once we had the key to access this door, the day came that we make computers in some way more "emotional."

4. Conclusion

Compared with single-language word embedding technology, the cross-lingual word embedding method mainly has the following two advantages: First, the cross-lingual word embedding method supports multi-language semantics, that is, reasoning about the meaning of language in multi-language contexts. It also supports calculating the similarity of cross-lingual corpora, closely related to many tasks (such as bilingual vocabulary induction or cross-lingual information retrieval). Second, the cross-lingual word embedding method supports the transfer of knowledge between different languages, especially between resource-rich and resource-poor languages. Meanwhile, it can be seen that the existing cross-lingual word embedding methods are based on the assumption that different languages

are semantically equivalent when describing the same thing. Then, by converting the cross-lingual knowledge transfer problem into a convex optimization problem, learning the mapping relationship between the semantic spaces of different languages to transfer the knowledge from labeled resource-rich languages to resource-poor languages.

References

[1] Lai, S., Liu, K., He, S., & Zhao, J. (2016). How to generate a good word embedding. *IEEE* Intelligent *Systems*, *31* (6), 5 - 14.

[2] Yin, Z., & Shen, Y. (2018). On the dimensionality of word embedding. *Advances in neural information processing systems*, *31*.

[3] Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C. C. J. (2019). Evaluating word embedding models: Methods and experimental results. *APSIPA transactions on signal and information processing*, *8*.

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

[5] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

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

[7] Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., & Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, *304*, 114135.

[8] Ruder, S., Vulić, I., & Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, *65*, 569 - 631.

[9] Zhao, J., Mukherjee, S., Hosseini, S., Chang, K. W., & Awadallah, A. H. (2020). Gender bias in multilingual embeddings and cross-lingual transfer. *arXiv preprint arXiv:2005.00699*.

[10] Mimno, D., & Thompson, L. (2017, January). The strange geometry of skip-gram with negative sampling. In *Empirical Methods in Natural Language Processing*.

[11] Dumais, S. T. (2004). Latent semantic analysis. Annual Review of Information Science and Technology (ARIST), 38, 189 - 230.

[12] Brown, D. E., & Huntley, C. L. (1992). A practical application of simulated annealing to clustering. *Pattern recognition*, *25* (4), 401 - 412.