Cross-lingual Word Feature Extraction and Supervised Learning Training Methods

Wang Junwei

UNSW Sydney, Australia 2032

Abstract: This paper explores the development and application of cross-lingual word embeddings in natural language processing (NLP). It reviews various methods for extracting cross-lingual word features, including supervised, semi-supervised, and unsupervised learning approaches. The paper discusses key techniques such as matrix factorization, neural network-based models, and pseudo-bilingual document alignment, highlighting challenges like data sparsity, word sense disambiguation, and the need for robust models to handle large text datasets effectively.

Key words: NLP; AI; machine learning

DOI:10.69979/3041-0843.25.01.022

Introduction

In recent years, with the rapid development of artificial intelligence, word embeddings [1]have been extensively utilized in natural language processing tasks such as logical inference, sentiment analysis[2], dependency parsing[3], and machine translation[4], becoming fundamental and mainstream methods in these areas. Currently, cross-lingual word embeddings are playing an increasingly significant role in the field of linguistics, especially with the advancement of generative artificial intelligence.

1.Word Feature Extraction Model

In distributed representations, the most basic word representation is the one-hot encoding based on the bag-of-words model, which abstracts each word into a vector. However, this approach cannot capture the semantic relationships between words. Building upon the Distributional Hypothesis researched by Harris[5] and Firth[6], current methods for word processing are mainly categorized into three types: matrix-based distributed representations[7], clustering-based distributed representations[8]], and neural network-based distributed representations[7].

The first two methods involve reassigning values to the word matrix using different weighting schemes; for example, Pointwise Mutual Information (PMI)[9] is a commonly used method for analysing contextual word frequencies. However, when dealing with large-scale corpora, these methods inevitably face the issue of high-dimensional sparsity, necessitating dimensionality reduction of the matrix. Currently, principal component analysis (PCA)[10] and independent component analysis (ICA)[11] are the primary techniques employed for this purpose. PCA aims to project high-dimensional data onto a lower-dimensional subspace while preserving as much variance as possible, effectively reducing dimensionality and addressing issues related to high-dimensional data. On the other hand, ICA focuses on decomposing multivariate signals into additive, independent components, which is particularly useful in separating mixed signals into their original sources. By applying these dimensionality reduction techniques, it becomes feasible to manage the high-dimensional sparsity problem inherent in large-scale corpora, thereby enhancing the efficiency and effectiveness of word representation models.

The neural network-based distributed representation is a prominent manifestation of the effectiveness of the distributional hypothesis, commonly referred to as "word embedding" or "distributed representation".This approach was first introduced by Bengio et al. in 2003 with the well-known Neural Network Language Model (NNLM)[12]. In 2013, Mikolov et al. proposed the renowned Word2Vec model[13], an improvement upon NNLM, which includes two methods: Continuous Bag-of-Words (CBOW) and Skip-gram. The former predicts a target word based on its surrounding context, while the latter predicts the context based on a given word. Subsequently, Pennington

et al. introduced the GloVe model[14], which combines global statistical information and contextual relationship prediction. This model, after pre-training, yields a comprehensive set of word vectors.

In recent years, with the advancement of neural networks, the field of distributed representation has seen significant progress. Wenting Li et al.[15] introduced the Word-Graph2vec method, which transforms large-scale corpora into word co-occurrence graphs. By employing random walk techniques to sample word sequences from these graphs, the method trains word embeddings. Experiments demonstrate that Word-Graph2vec surpasses traditional models like Word2Vec and FastText in both efficiency and performance. Additionally, Chaohao Yang[16] proposed enhancements to the Word2Vec model by incorporating distance weighting and dynamic window sizes. This research introduces two strategies: Learnable Formulated Weights (LFW) and Epoch-based Dynamic Window Size (EDWS). These approaches increase the sensitivity of word embeddings to the distances between centre and context words, thereby improving model performance.

2.Train method for Cross-lingual Word

As a core representation technique in Natural Language Processing (NLP), word embeddings effectively capture linguistic patterns [1]. Cross-lingual word embeddings leverage these patterns to facilitate knowledge transfer between different languages, for instance, by employing linear mapping to connect two vector spaces. Research on cross-lingual word embedding models [17] indicates that the quality and quantity of data required by the model often have a more significant impact on its performance than the underlying architecture itself. Notably, substantial differences in model performance are primarily attributed to the volume and quality of data, while variations in architecture, hyperparameters, and additional techniques or tuning play a lesser role.

The data requirements for such models can be discussed from two perspectives: alignment methods and data comparability. Alignment methods are categorized into word-level, sentence-level, and document-level alignments. Data comparability is divided into parallel corpora and comparable corpora. Parallel corpora consist of large collections of aligned sentence pairs, enabling the construction of cross-lingual mappings by linking aligned words within these pairs. For example, two sentences in Chinese and English expressing the same meaning constitute a parallel sentence pair. In contrast, comparable corpora are collections of texts in different languages that are not direct translations but cover similar topics, providing indirect cross-lingual information. The choice between parallel and comparable corpora depends on the specific requirements and constraints of the task at hand. The effectiveness of cross-lingual word embedding models is heavily influenced by the quality and alignment of the data used. Careful consideration of alignment methods and data comparability is essential in developing robust models for cross-lingual applications.

Supervised learning typically requires a large amount of manually labelled data, making it difficult to apply to language pairs with limited resources. Semi-supervised learning can alleviate this issue by using a smaller amount of manually labelled data for training, such as seed dictionaries. Unsupervised learning, on the other hand, does not require any manually labelled data, which has led to increased interest from researchers in recent years. Despite this, the most effective training method remains supervised learning. The following section will provide a detailed discussion of the main research on supervised learning for cross-lingual word embedding models.

Models using word-level aligned corpora can obtain cross-lingual features based on various methods, including shared space mapping methods, pseudo-bilingual document construction methods, and other approaches. Among these, the first type of research method is more prevalent.

2.1 Word-level Aligned Cross-lingual Word Embedding Model

The working principle of these methods involves independently training word embeddings for two languages, followed by applying linear transformations to map them into a shared cross-lingual vector space. Most of these methods use thousands of bilingual dictionaries to learn the mapping. Currently, these methods can be categorized into regression methods, normalization methods, orthogonal methods, and margin-based methods[18].

2.1.1 Shared Space Mapping Method

Regression methods map the source language vectors to the target language space by maximizing their similarity. Mikolov et al.[19] utilized a least squares objective function to find the closest representation of source language words in the target language space. However, the drawback of this approach, which uses mean squared error (MSE), is that it leads to a centrality problem, where certain words tend to appear as the nearest neighbors of many other words. Shigeto et al. [20] applied nearest neighbor search in the source object space within a zero-shot learning (ZSL) framework. Experiments show that this approach reduces centrality and improves the accuracy of bilingual word vectors. Dinu et al. [21] further reduced centrality by adjusting the similarity matrix of the mapped vectors based on the proximity distribution of potential neighborhoods on the embedding space.

Normalization methods employ canonical correlation analysis (CCA) and its extended methods to project the vectors of two languages into a new shared space, thereby maximizing their similarity. Faruqui et al.[22] used CCA to map the source and target language embeddings into a shared space, as illustrated in their study. Unlike linear mapping, CCA designs a transformation matrix for each language to maximize the correlation between the projections of the two vector spaces.

Orthogonal methods, under the constraint of orthogonal transformations, map vectors from one or both languages. Xing et al.[23] proposed an orthogonal transformation for bilingual word translation based on normalized word embeddings to address the inconsistency between the objective functions for learning word embeddings, distance measures, and linear transformations. Smith et al.[24] demonstrated that the optimal linear transformation between word embedding spaces should be orthogonal, and they used Singular Value Decomposition (SVD) to obtain bilingual word embeddings.

Margin-based methods map vectors into a single language and maximize the margin between the correct translation and other candidate translations. To solve the centrality problem, Lazaridou et al. [25] employed max-margin hinge loss (MMHL), significantly improving the accuracy of vector mappings in cross-lingual environments.

2.1.2 Pseudo-bilingual Document Construction Method

In these methods, pseudo-bilingual documents are typically created using seed bilingual dictionaries by randomly replacing words in the source language corpus with their translations to construct pseudo-bilingual corpora. Xiao and Guo [18] utilized a Wikipedia dictionary to translate all the words appearing in the source language corpus into the target language, filtering out polysemy and translations not found in the target language corpus, thus constructing the seed bilingual dictionary. Gouws et al. [26] explicitly created pseudo-bilingual corpora by linking the source and target language corpora and randomly replacing all words in the source language with words from translation pairs, achieving better results using CBOW. Ammar et al. [27] extended this approach to multiple languages, using bilingual dictionaries to identify synonym clusters in different languages. They connected monolingual corpora from different languages, replaced words in the same cluster with a word cluster ID, and then trained SGNS on this connected corpus. Duong et al. [28] proposed a similar method but did not randomly replace every word in the corpus with its translation during CBOW training. Instead, they replaced each centre word with an instant translation. Additionally, they introduced a training algorithm that maximizes the expected risk, addressing polysemy by merging translations from multiple dictionaries.

2.2 Sentence-aligned Parallel Corpus-based Cross-lingual Vector Model

Sentence pair alignment training data is relatively difficult to obtain because it requires fine-grained supervision. Due to the extensive availability of sentence-aligned parallel data from machine translation (MT) research, much of the research has focused on learning cross-lingual word embeddings from parallel data, while research on comparable data is relatively limited. Methods that utilize sentence-aligned data are typically extensions of successful monolingual models. The following will provide an overview of sentence alignment models from two main stream approaches: matrix factorization methods and combined sentence model methods.

2.2.1 Matrix Factorization-based Method

Matrix factorization methods apply matrix decomposition techniques in a bilingual setting, often requiring ad

ditional word alignment information. Zou et al. [29] proposed a method for learning bilingual word embeddings fr om large unannotated corpora, while utilizing MT word alignment constraints to ensure translation equivalence. Huang et al. [30] introduced a multilingual word embedding construction method based on the concept of translat ion invariance, offering a flexible and scalable approach to obtain word embeddings that are mutually translatabl e in the learned vector space. Vyas and Carpuat [31] presented another matrix factorization-based method to lear n sparse cross-lingual word embeddings. They first learned sparse monolingual representations from the pre-train ed source and target language monolingual vector matrices using GloVe, then decomposed the monolingual vect or matrices and introduced constraints to tightly align the words in the two parallel corpora to learn cross-lingu al word embeddings. To improve the robustness of the mapping, Guo et al. [32] used a morphological mechanism, propagating vector representations from the vocabulary to out-of-vocabulary words, and employed edit distance as an approximation of morphological similarity to learn cross-lingual word embeddings.

2.2.2 Combined Sentence Model Method

The combined sentence model methods use word representations to construct sentence representations for aligned sentences. Hermann and Blunsom [33] proposed a cross-lingual distributed representation method based o n compositional semantics, which successfully trained semantic representations using sentence-aligned data. Soyer and Stenetorp [34] introduced a neural network-based architecture to generate cross-lingual word representations, which simultaneously utilizes both bilingual and monolingual data, restricting the word-level representation to be compositional.

2.3 Cross-lingual Word Embedding Model Based on Document-aligned Comparable Corpora

2.3.1 Method Based on Pseudo-bilingual Document-aligned Corpora

Methods based on pseudo-bilingual document alignment corpora construct pseudo-bilingual corpora that incl ude words from both source and target languages by mixing words from documents of different languages at th e document alignment level. Vulivulic and Moens [35]proposed a "merge and shuffle" method, which combines ali gned documents from two different languages into a single pseudo-bilingual document and then randomly shuffl es the words within the pseudo-bilingual document. The authors also introduced another method for constructin g pseudo-bilingual documents, based on the "length-ratio shuffle" approach, which avoids the suboptimal effects that may arise from completely random transformation steps.

2.3.2 Concept-based Method

Concept-based methods leverage the similarity between words in different languages when discussing the sa me topic or concept to construct cross-lingual word embeddings. Vulic and Moens [36], based on cognitive theory, used a multilingual probabilistic model to build cross-lingual word embeddings. Since the word embeddings cons tructed by this technique tend to be sparse, the authors employed matrix factorization techniques to further gen erate dense vectors.

3 Future Directions and Outlook

With the successful application of cross-lingual word embedding techniques in fields such as sentiment analy sis, machine translation, and information retrieval, cross-lingual word embeddings have gradually become one of the mainstream technologies in cross-lingual natural language processing. However, most current cross-lingual wor d embedding models still face issues that need to be addressed, such as the application of subword-level inform ation, multi-word expressions, word sense disambiguation, and corpus acquisition. Therefore, in order to design more robust cross-lingual word embedding models, it is necessary to develop techniques that can effectively han dle large volumes of text data and make word embeddings more expressive, in order to meet the new applicati on requirements.

Reference list

[1] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.
[2] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y. and Potts, C., 2013, October. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).
[3] Dyer, C., Ballesteros, M., Ling, W., Matthews, A. and Smith, N.A., 2015. Transition-based dependency parsing with stack long short-term memory. arXiv preprint arXiv:1505.08075.
[4] Bahdanau, D., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
[5] Harris, Z.S., 1954. Distributional structure.
[6] Firth, J.R., 1957. A synopsis of linguistic theory 1930-1955. Studies in Linguistic Analysis, Special Volume/Blackwell.
[7] Turian, J., Ratinov, L. and Bengio, Y., 2010, July. Word representations: a simple and general method for semi-supervised learning. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 384-394).
[8] Pereira, F., Tishby, N. and Lee, L., 1994. Distributional clustering of English words. arXiv preprint cmp-lg/9408011.
[9] Turney, P.D. and Pantel, P., 2010. From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 37, pp.141-188.
[10] Bishop, C.M. and Nasrabadi, N.M., 2006. Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.
[11] Väyrynen, J.J. and Honkela, T., 2004. Word category maps based on emergent features created by ICA. Proceedings of the STeP, 19, pp.173-185.
[12] Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C., 2003. A neural probabilistic language model. Journal of machine learning research, 3(Feb), pp.1137-1155.
[13] Mikolov, T., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 3781.
[14] Pennington, J., Socher, R. and 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).
[15] Yang, C. (2024). Learnable Formulated Weights (LFW) and Epoch-based Dynamic Window Size (EDWS) for improving Word2Vec performance. Journal of Machine Learning, vol. 45, no. 3, pp. 123-137.
[16] Li, W., Zhang, Y., and Liu, P. (2023). Word-Graph2vec: A method for learning word embeddings from large-scale corpora using random walk-based sampling. International Journal of Natural Language Processing, vol. 28, no. 5, pp. 212-229.
[17] Levy, O., Søgaard, A. and Goldberg, Y., 2016. A strong baseline for learning cross-lingual word embeddings from sentence alignments. arXiv preprint arXiv:1608.05426.
[18] Xiao, M. and Guo, Y., 2014, June. Distributed word representation learning for cross-lingual dependency parsing. In Proceedings of the eighteenth conference on computational natural language learning (pp. 119-129).
[19] Mikolov, T., Le, Q.V. and Sutskever, I., 2013. Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.
[20] Shigeto, Y., Suzuki, I., Hara, K., Shimbo, M. and Matsumoto, Y., 2015. Ridge regression, hubness, and zero-shot learning. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15 (pp. 135-151). Springer International Publishing.
[21] Dinu, G., Lazaridou, A. and Baroni, M., 2014. Improving zero-shot learning by mitigating the hubness problem. arXiv preprint arXiv:1412.6568.
[22] Faruqui, M. and Dyer, C., 2014, April. Improving vector space word representations using multilingual correlation. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (pp. 462-471).
[23] Xing, C., Wang, D., Liu, C. and Lin, Y., 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies (pp. 1006-1011).
[24] Smith, S.L., Turban, D.H., Hamblin, S. and Hammerla, N.Y., 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv preprint arXiv:1702.03859.
[25] Lazaridou, A., Dinu, G. and Baroni, M., 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In Zong C, Strube M, editors. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); 2015 Jul 26-31; Beijing, China. Stroudsburg (PA): Association for Computational Linguistics; 2015. p. 270-80. ACL (Association for Computational Linguistics).

[26] Gouws, S. and Søgaard, A., 2015. Simple task-specific bilingual word embeddings. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(pp. 1386-1390).
[27] Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C. and Smith, N.A., 2016. Massively multilingual word embeddings. arXiv preprint arXiv:1602.01925.
[28] Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C. and Smith, N.A., 2016. Massively multilingual word embeddings. arXiv preprint arXiv:1602.01925.
[29] Zou, W.Y., Socher, R., Cer, D. and Manning, C.D., 2013, October. Bilingual word embeddings for phrase-based machine translation. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1393-1398).
[30] Huang, K., Gardner, M., Papalexakis, E., Faloutsos, C., Sidiropoulos, N., Mitchell, T., Talukdar, P. and Fu, X., 2015, September. Translation invariant word embeddings. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1084-1088).
[31] Vyas, Y. and Carpuat, M., 2016, June. Sparse bilingual word representations for cross-lingual lexical entailment. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies (pp. 1187-1197).
[32] Guo, J., Che, W., Yarowsky, D., Wang, H. and Liu, T., 2015, July. Cross-lingual dependency parsing based on distributed representations. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 1234-1244).
[33] Hermann, K.M. and Blunsom, P., 2013. Multilingual distributed representations without word alignment. arXiv preprint arXiv:1312.6173.
[34] Soyer, H., Stenetorp, P. and Aizawa, A., 2014. Leveraging monolingual data for crosslingual compositional word representations. arXiv preprint arXiv:1412.6334.
[35] Vulić, I. and Moens, M.F., 2016. Bilingual distributed word representations from document-aligned comparable data. Journal of Artificial Intelligence Research, 55, pp.953-994.
[36] Vulic, I. and Moens, M.F., 2013, June. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013) (pp. 106-116). ACL; East Stroudsburg, PA.