Plongement multilingue non supervisée basé sur le mapping monolingue pour la traduction automatique neuronale des langues faiblement dotées

Martin Rodrigue Atangana Ongolo, Paulin Melatagia Yonta

Unsupervised multilingual word embedding based on monolingual mapping for neural machine translation of low resources languages

Martin Rodrigue ATANGANA ONGOLO[1] **and Paulin MELATAGIA YONTA**[1,2]

[1]Department of Computer Science, University of Yaoundé I, Cameroon
[2]IRD, UMMISCO, F-93143, Bondy, France

*E-mail : atanganamr@yahoo.com, paulinyonta@gmail.com

Abstract

Limited data resources are the current problem for neural machine translation of low resources languages. In this paper we propose to use multilingual embedding as a method of representing words as input to a neural machine translator. This approach is then compared to word representation approaches in monolingual context used in the literature for neural machine translation. The results with multilingual embedding on a dataset of 7187 pairs of French-Ewondo parallel verses of the bible are promising because they are better than those obtained with the representation approaches used so far. We obtained 8.77% of blue for the unsupervised multilingual embedding, 5.34% for the semi-supervised version and finally 4.92% for the word embedding with the skip-gram architecture.

Keywords

Multilingual word embedding, Neural networks, Neural machine translation, low resources languages.

I INTRODUCTION

Machine translation is a natural language processing task that consists of transcribing an original text written in a source language into a target language in which the translation is desired. Two main approaches are currently used : the statistical approach and more recently the neural approach[4]. The statistical approach generates a translated document based on statistical models obtained through the analysis of multilingual databases and the neural approach relies on artificial neural networks to learn linguistic rules from the statistical models. The translations obtained with the neural approach are better in terms of quality[1]. It is crucial to represent these words in a form that can be manipulated by neural networks models while preserving their semantics as much as possible and in a relevant way in order to produce quality translations. This is how onehot encoding and word embedding methods are used as word representation techniques for neural machine translation. Unfortunately, the onehot encoding produces sparse vectors (full of zeros) which do not encapsulate the relevant semantic distance between words and moreover these vectors are expensive in terms of memory resources. The second word representation approach used is based on word embedding methods, in particular the word2vec architectures proposed in 2013 by Mikolov et al.[13]. These methods produce dense word representation vectors, encapsulate the semantic relationships between words of one language but do not capture the semantic relationships between words of different languages. In particular,

the small size of the corpora of poorly endowed languages poses a performance problem for monolingual learning because monolingual methods need a large amount of data to train well and produce good word representation models[5].

In this article, we propose to use multilingual word embedding based on monolingual mapping, which consists of constructing the source and target word representation vectors in the same space so that the words of the source and target languages with similar meanings have close representation vectors to improve neural machine translation for French - Ewondo. Ewondo is a language spoken in the central part of Cameroon by more than $2,500,000$ people. It is the most widely spoken national language in Yaounde and neighboring regions.

The rest of the paper is organized as follows : Section 2 presents word representation for neural machine translation, Section 3 presents multilingual embedding based on monolingual mapping and Section 4 presents experiments and discussions. Section 5 concludes the paper.

II WORD REPRESENTATION FOR NEURAL MACHINE TRANSLATION

Current neural machine translation architectures are based on the encoder-decoder structure proposed in 2013 by Kalchbrenner et al.[10]. Both encoder and decoder are neural networks. The encoder takes as input the source sentence to be translated which is here a sequence of source words $X_1, X_2....X_n$ and each source word is represented by its representation vector. It will transform the input sequence of words into a context vector V_c of fixed size. The decoder will then take this context vector as input and generate as output the sequence of target words $S_1, S_2, ..., S_m$ such that the probability $P(S_1, S_2, ..., S_m|X_1, X_2....X_n)$ is maximum. There are three neural network architectures in neural machine translation : the first one based on convolutional neural networks, the second one based on recursive neural networks and the third one based on attention mechanisms[12].

In the architecture of recurrent neural networks, the encoder and the decoder are both recurrent neural networks. The encoder takes as input a sequence of words $W = (W_1, ..., W_n)$ represented by the word representation vectors $X = (X_1, ..., X_n)$. The model parameters are $\theta = (U, V, W)$ where U is the weight matrix located between the cells of the hidden layer h_i and h_{i-1}, V the matrix of weight located between the hidden cells h_i and the output cells y_i, W the matrix of weights located between the input cells X_i and the hidden cells h_i.

In the architecture based on convolutional neural networks, the encoder and the decoder are both convolutional neural networks. The encoder takes as input a sequence of n words $W = (W_1, ..., W_n)$ represented by the representation vectors $X = (X_1, ..., X_n)$. It apply a 1D(1-dimension) convolution with a filter and then obtain n vectors of m-dimensions $Z = (Z_1, ..., Z_n)$. The vectors Z_i are sent to the level of the hidden layer H composed of the GLU (Gated Linear Unit) cells H_i[7].

The architecture based on the attention mechanism was proposed in 2015 by Bahdanau et al.[3]. It is a solution to the problem of the fixed size context vector generated by the encoder. Instead of encoding the input sequence into a single vector fixed-size context, the idea here is to generate a context vector C_t at each time step t output from encoder[3].

2.1 The onehot encoding

Given a vocabulary $V = \{w_1, w_2, ..., w_n\}$ the onehot encoding consists in representing a word w by a vector T of size $|V|$ such that $T[i] = 1$ if $w = w_i$ and $T[i] = 0$ otherwise. The onehot

encoding produces sparse vectors (full of zeros) and the vectors obtained do not describe any notion of similarity between the words. To overcome these limits, word embedding methods have emerged.

2.2 Word embedding methods

Word embedding methods can be grouped into three main families : methods based on the local context window, methods based on matrix factorization and hybrid methods.

2.2.1 The methods based on the local context window

Word2vec is a model proposed in 2013 by Mikolov et al.[13]. It is based on a two-layer neural network, and makes it possible to learn the vector representations of words composing a text corpus, so that the words sharing similar contexts are represented by close numerical vectors. This model presents two architectures, namely : the CBOW (Continuous Bag Of Words) architecture which consists of predict a central word w given its input context $c = \{w_1, w_2, ..., w_n\}$ and the SKIPGRAM architecture which does the inverse of CBOW[13]. Skipgram Negative Sample is a variation of Skipgram that involves downsampling the words into two sets : the set D^+ made up of pairs of words (w_i, w_j) such that w_i and w_j appear together in the text corpus and the set D^- consisting of pairs of words (w_i, w_k) such that w_i and w_k do not appear together in the corpus of text[8].

2.2.2 Methods based on the factorization of matrices

These methods are based on two main concepts : construction of the co-occurrence matrix and dimension reduction. The Latent Semantic Analysis (LSA) method is an unsupervised method of learning vector representations of words proposed in 1990 by Scott Deerwester and al[6]. It works as follows :

— The first step consists in constructing a co-occurrence matrix of terms or words $A = [a_{ij}]$ of size $t \times d$ with t the number of words and d the number of documents, a_{ij} is the frequency or importance of the word i in the document j.
— The second step consists in identifying the singular values of the matrix A in order to be able to decompose it into three matrices $T0$, $S0$, $D0$ such that $A = T0.S0.D0^T$.
— The third step consists in reducing the diagonal matrix $S0$ of size $m \times m$ to a diagonal matrix S of size $k \times k$ with $k \leq m$.
— The fourth step consists in deducing the matrix A' of size $t \times d$, the approximation at rank k of the matrix A is $A' = T.S.D^T$ where T is a (t, d) matrix of d documents and t terms and D is a unitary matrix. The matrix A' thus obtained is an approximate matrix of the initial matrix A.

2.2.3 Hybrid methods

Hybrid methods are based both on the principle of local context window models and matrix factorization models. We can cite among these methods : the Glove method which is a monolingual method vector representation of unsupervised learning words proposed in 2014 by Pennington et al.[11]. It is based on the construction of a co-occurrence matrix M of words, using a sliding popup over the text corpus.

In the case of low resources languages, the size of the corpora is small, which raises a performance problem in learning for monolingual methods because monolingual methods need a

large amount of data to train well and produce good word representation models. Also the one-hot and word embedding methods capture only the semantic relationships of words in a single language, we propose the use of the word embedding method based on monolingual mapping as a word representation method for translation. neural automation to benefit not only from the volume of data in the well-endowed languages but also to capture the semantic relationships that exist between source and target languages to enrich word representation vectors.

III MULTILINGUAL EMBEDDING BASED ON MONOLINGUAL MAPPING

This method is inspired by multilingual word embedding approaches, in particular the work of Mikolov et al.[9]. Proposed in 2013, it consists of constructing the source and target word representation vectors by sharing in the same space via a linear transformation matrix the source and target word representation vectors resulting from the monolingual word embedding method such that source and target language words with similar meanings have similar representation vectors. The word embedding technique used is SkipGram Negative Sample. Figure 1 shows how it works.

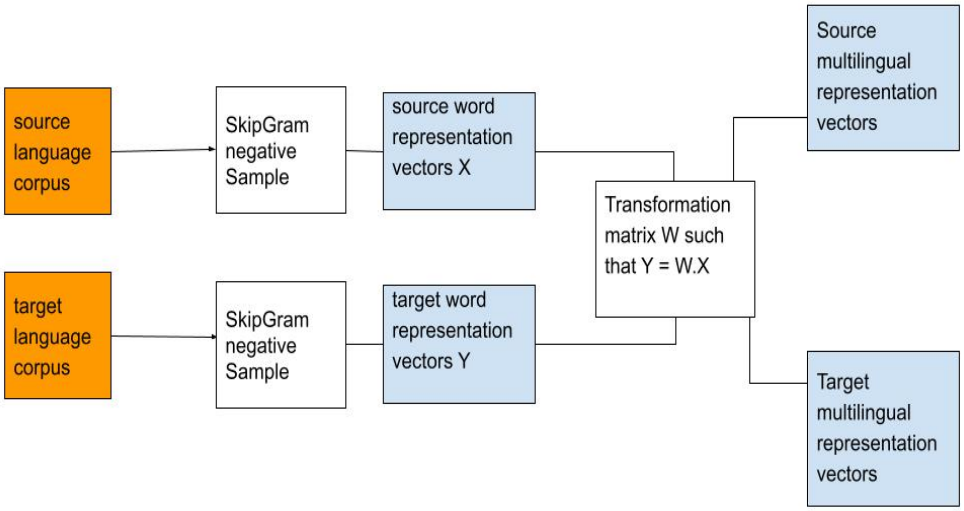

FIGURE 1 – Multilingual embedding based on monolingual mapping

Given the source C_s and target C_t corpora, the source and target representation vectors are learned using the SkipGram Negative Sample word embedding technique. The obtained source X and target Y representation vectors are used to construct a transformation matrix W such that $Y \approx W.X$. Once the transformation matrix W has been obtained, the source word representation vectors enriched Z_i by multilingual embedding are obtained on the one hand from the source monolingual vectors X_i and from the transformation matrix W by the equation :

$$Z_i = X_i.W \tag{1}$$

On the other hand, the enriched target word representation vectors S_i by multilingual embedding are obtained from the target monolingual vectors Y_i and from the transformation matrix W by the equation :

$$S_i = Y_i.W^{-1} \tag{2}$$

The main property of this solution is inspired by the work of Chao Xing et al.[15]. In order to avoid loss or degradation of monolingual semantic information when calculating the product of matrices between monolingual vectors and the transformation matrix W, we constrain the transformation matrix W to be orthogonal : $I_n = T_W.W$ with $T_W = W^{-1}$

Depending on how to obtain the transformation matrix W, there are three variants of the solution : the unsupervised version, the semi-supervised version and the supervised version. We denote by X_s the matrix containing the monolingual representation vectors of the source words and by X_t the matrix containing the monolingual representation vectors of the target words.
In the unsupervised version, the transformation matrix W is calculated analytically by the Moore-Penrose pseudo-Inverse method by the equation :

$$W = X^{+}.X_t \tag{3}$$

with

$$X^{+} = (X_s^T.X_s)^{-1}.X_s^T \tag{4}$$

For the supervised version, the transformation matrix W on a bilingual dictionary D is learned such as :

$$D_i.W \approx D_j \tag{5}$$

Where D_i represents the source words and D_j their target translations. Finally in the semi-supervised version the transformation matrix is initially calculated by the Moore-Penrose pseudo-inverse method then we learn W on a very small bilingual dictionary[14] .

IV EXPERIMENTS

4.1 Experimental Protocol

The preparation of the data consisted of using scrapping to extract the parallel Bible verses (French - Ewondo). From these biblical verses, we build a dataset of 7187 pairs of French-Ewondo parallel sentences taken from the bible in the ewondo edition2012 language, in particular from the New Testament. Then we performed a schuffle to mix the random way of these sentences. and finally subdivide the dataset as shown in Table 1.

Training(Train set)	Validation(Dev set)	Evaluation(Test set)
5187	1000	1000

TABLE 1 – Dataset repartition

In practice the onehot encoding is difficult to implement in our case since we had a source corpus vocabulary of 14000 words this means that we would have to build 14000 word representation vectors where each vector would be of size 14000 which is very costly in terms of memory resources. To do multilingual embedding, we used the vec2map framework proposed in 2016 by Mikel Artetxe et al.[2]. We experimented with the semi-supervised version (using a

dictionary of 25 words) and non-supervised version. The supervised version is difficult to implement because not only are word dictionaries very rare for low resources languages, but also building a dictionary of more than one hundred words is difficult. We used BPE (Byte Pair Encoding) as a tokenization technique. With the Fairseq tool, we used the *transformer_iwslt_de_en* architecture which is an architecture used in the Fairseq documentation for the automatic translation from German to English on 555 epochs whose particuliarity is that its hidden layer has a size of 1024 instead of 2048 on the base fairseq transformer architecture. The choice of this architecture is also due to the intuition that as we are in a weakly endowed environment, less complex neural models may be adequate.

4.2 Results and discussions

Skipgram, semi-supervised and unsupervised multilingual embedding representation approaches were evaluated on the Blue measurement over 555 epochs, the following table presents the results obtained :

Representation approaches	Score Blue(%)
Skipgram monolingual method	4.92
Semi-supervised multilingual Based on monolingual mapping	5.34
Unsupervised multilingual Based on monolingual mapping	**8.77**

TABLE 2 – Results of different word representation methods for machine translation on our dataset.

These results show the word representation vectors resulting from multilingual word representation methods based on monolingual mapping, particularly the unsupervised version, give good performance compared to the Skipgram monolingual representation method for the machine translation of Ewondo using our dataset.

V CONCLUSION

Neural machine translation of low-resource languages is getting a lot of attention in natural language processing. Works proposed in the literature use onehot encoding and monolingual word embedding methods to construct representation vectors. In this work we proposed to use multilingual embedding based on monolingual mapping as a word representation method to improve neural machine translation. Experiments carried out on a dataset of 7187 pairs of French-Ewondo parallel verses of the Bible show us that this approach gives better performance than those based on the skipgram word embedding method on the blue measure. While the obtained bleu score is low, this approach seems promising. In perspective, we propose to carry out experiments on another language and to increase the size of the corpus in order to obtain exploitable models.

REFERENCES

[1] Ankush and M. Agarwal. *Machine translation : a literature review, arXiv preprint arXiv :1901.01122*. 2018.

[2] Artetxe, G. Labaka, and E. Agirre. *Learning principled bilingual mappings of word embeddings while preserving monolingual invariance, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016.

[3] Bahdanau, K. Cho, and Y. Bengio. *Neural Machine Translation by Jointly Learning to Align and Translatey, arXiv preprint arXiv :1409.0473*. 2014.

[4] Beyala and M. Nkenlifack. *Extended convolutional neural networks post-trained with factored statistical machine, International Research Journal of Computer Science,197-208*. Aug. 2020.

[5] Bojar, C. Buck, C. Callison-Burch, B. Haddow, P. Koehn, C. Monz, M. Post, H. Saint-Amand, R. Soricut, and L. Specia. « Proceedings of the Eighth Workshop on Statistical Machine Translation ». In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. 2013.

[6] Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. *Indexing by latent semantic analysis, Journal of the American society for information science*. 1990.

[7] Gehring, A. Michael, D. Grangier, D. Yarats, and Y. N. Dauphin. *Convolutional sequence to sequence learning, International Conference on Machine Learning, 1243–1252*. 2017.

[8] Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. *Distributed Representations of Words and Phrases and their Compositionality, Advances in Neural Information Processing Systems 26*. 2013.

[9] Mikolov, L. Q. V, and I. Sutskever. *Exploiting similarities among languages for machine translation, arXiv preprint arXiv :1309.4168*. 2013.

[10] K. Nal and P. Blunsom. *Recurrent continuous translation models, Proceedings of the 2013 conference on empirical methods in natural language processing*. 2013.

[11] Pennington, R. Socher, and C. D. Manning. *Glove : Global vectors for word representation, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

[12] Tang, M. Müller, A. Rios, and R. Sennrich. *Why self-attention ? a targeted evaluation of neural machine translation architectures, arXiv preprint arXiv :1808.08946*. 2018.

[13] Tomas, K. Chen, G. Corrado, and J. Dean. *Efficient Estimation of Word Representations in Vector Space, Computational Linguistics*. 2013.

[14] Vulić, G. Glavaš, R. Reichart, and A. Korhonen. *Do we really need fully unsupervised cross-lingual embeddings ?, arXiv preprint arXiv :1909.01638*. 2019.

[15] Xing, W. Dong, C. Liu, and Y. Lin. *Normalized word embedding and orthogonal transform for bilingual word translation, Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. 2015.