2016 International Conference on Artificial Intelligence: Techniques and Applications (AITA 2016)
ISBN: 978-1-60595-389-2

Research on Deep Learning HMM Word Alignment

Dan LI[1] and Zheng-hong YU[2,*]

[1]The College of Post and Telecommunication of WIT, Wuhan, China

[2]Wuhan University of Science & Technology City College, Wuhan, China

*Corresponding author

Keywords: Deep learning, Word Alignment, HMM.

Abstract. Word alignment is an essential step for machine translation. According to the over-fitting of traditional word alignment model and the weakness for the context description, the deep learning hidden HMM word alignment combined a multi-layer neural network with an undirected probabilistic graph model, use the similarity of the word and context information word alignment to be a more precise modeling. Experimental results show that, compared with the reference system, this model can significantly improve the effect of word alignment, which is applicable.

Introduction

Machine translation contains three major issues, which are word alignment, ordering and translation modeling. Among them, word alignment is the basic problem for machine translation, and is also the first step for today's mainstream machine translation. By using word alignment, machine translation system could automatically find the lexical level correspondences from the bilingual training corpus of the sentence alignment. Then according to the results of word alignment, we have further extraction based on translation rules with the heuristic method. Besides, we could train the machine translation model to increase the accuracy, to improve the performance of the machine translation system.

Traditional word alignment model used sparse lexical features, which were less effective at the low-frequency word alignment in the bilingual database. Besides, Traditional word alignment model had the weak ability for describing the context, which was unable to effectively describe the similar word alignment based on context model.

Machine translation researchers did a large amount of studying to improve the effect of word alignment. Wu[1] proposed inverted transduction grammar, which was trying to limit the search space of word alignment model by using linguistic formal syntax. Deng[2] extended original lexical HMM to phrasal HMM, to improve the one-to-many alignment description. He[3] used lexical reordering model to improve the weakness of the reordering description. Liang[4] had joint optimization for one language on word alignment on both direction. Berger[5] initially gave out the word alignment based on maximum entropy. Moore[6] put forward the word alignment based an informal log-linear model. All the above methods, which used supervised training method, limited its feature selection because there were limited word alignment database based on manual annotation.

Deep Learning Model

Deep learning achieves the approximation of complex function, characterization of the input data by learning a deep nonlinear network, and shows the great power in extracting the intrinsic feature of the training data.

Compared with the surface structure, the deep structure uses less parameter to reflect more complex mapping relationship, which is proved to be a more compact representation. Deep learning, with the feature of learning distributional representation automatically, could have powerful global generalization performance. Meanwhile, deep learning with the feature of the hierarchical learning, would reduce workload of artificial feature and produce newly useful feature. Compared with the

breakthrough in the field of the sound and the image, deep learning in the field of natural language progresses is still at the stage of researching.

Traditional HMM(Hidden Markov Model)

Word alignment aims at finding the lexical corresponding relation from bilingual corpus between different languages, especially for sentences. Different from the translation method of word alignment model, which is referred as a generation probability from one language to another, word alignment is the hidden variable in this generation probability model.

HMM word alignment also treats word alignment as a hidden variable for the translation process from the original language sentence to target language sentence. And HMM word alignment not only offer an effective accurate deduction, but also provides better alignment result.

$$p(a_i \mid a_1^{i-1}, f_1^{i-1}, m, e) \approx (pa_i - a_{i-1}, m)$$

$$p(f_i \mid a_1^{i}, f_1^{i-1}, m, e) \approx p(f_i \mid e_{a_i}) \tag{1}$$

$$p(f, a \mid e) = \prod_{i-1}^{n} p(f_i \mid e_{a_i}) p(a_i - a_{i-1}, m)$$

Among the Eq.1, f corresponds to the observation-sequences of HMM and a refers to the hidden states of HMM model, while $p(f_i \mid e_{a_i})$ corresponds to the emission probability and $p(a_i - a_{i-1}, m)$ refers to state transition probability of the HMM model. The parameter of HMM includes two parts, one is the vocabulary translation probability tables $p(f_i \mid e_{a_i})$, which is describing the corresponding relation of the context-free vocabulary; another one is the order probability tables based on the skip distances $p(a_i - a_{i-1}, m)$, which is describing the different word order between different languages.

Traditional HMM word alignment model has two limitations. Firstly, HMM model couldn't have an accurate evaluation for the low-frequency vocabulary, which would cause the over-fitting and result in a garbage collector made by the low-frequency vocabulary. Secondly, HMM word alignment model has weaker description ability for the context, which couldn't reflect the complicated language phenomenon for the description of two different word orders, just by simple skip distances.

Deep learning HMM word alignment model based on the context

This paper presents context-dependent deep learning HMM word alignment (CDDL-HMM), which is a word alignment model based on deep learning neural network. It combines neural network with the traditional HMM word alignment, uses the neural network to reduce the dimension for the sparse lexical feature, and describes the similarity of the words and the contextual relevance to gain a better word alignment result.

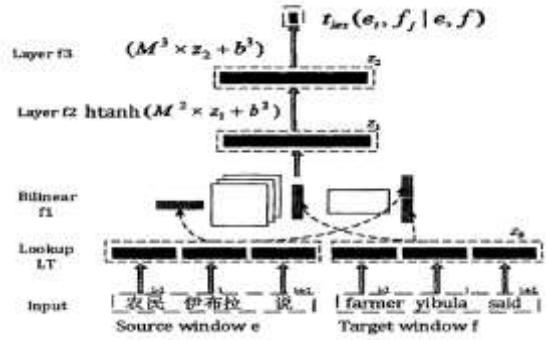

Figure 1.Structure of word alignment bi-linear neural network.

Deep learning neural network model is an extension of the traditional HMM word alignment model. We mark the lexical translation probability table of traditional HMM as p_{lex}, and label the skip probability table as p_d. Fora bilingual sentence pair (e,f) and one of its alignments a, HMM provides its given probability for the triple (e,f,a), as following Eq.2.

$$p(f,a|e)=\prod_{i-1}^{m} p(f_i|e_{a_i})\,p(a_i-a_{i-1},m) \tag{2}$$

In order to combine the deep neural network and the traditional HMM word alignment, we transfer the directed probabilistic graphic of HMM into an undirected probabilistic graphic, and calculate its potential function with a deep neural network. More specifically, for a given bilingual sentence pair, the neural network model would provide its probability as following Eq.3.

$$p(f|e)=\frac{\sum_a \exp\psi(e,f,a)}{\sum_{f'}\sum_a \exp\psi(e,f',a)},$$

$$p(f,a^*|e)=\frac{\exp\psi(e,f,a^*)}{\sum_a \exp\psi(e,f',a)} \tag{3}$$

In the above calculation formulas, Ψ is a potential function in sentence level. And it will be further divided into.

$$\psi(e,f,a)=\sum_{j=1}^{|f|}\phi(f_j,a_j,e)+\varphi(a_{j-1},a_j,e) \tag{4}$$

In the above calculation formulas, ø is translation component and ψ is jump component. This kind of decomposition, essentially, is a method of decomposing an undirected graphic into a conditional random field (CRF, for short). Usually, the potential function of CRF is a linear function, which is reflected by F, its feature vectors, and the inner product form of its corresponding feature weight vector W.

$$\phi(\cdot)=F_\phi(\cdot)\cdot w_\phi,\ \psi(\cdot)=F_\psi(\cdot)\cdot w_\psi \tag{5}$$

Then, the potential function of the corresponding relation is modified as following.

$$\phi(f_j,a_j,e)=F_\phi(f_j,a_j,e)\cdot w_\phi+S_\phi(f_j,a_j,e) \tag{6}$$

In the above calculation formulas, S* refers to a multi-layer neural network. This neural network takes the source language and the target language f_j in the context-window as input. It finally outputs a real as a component for the e_i and f_j alignment.

$$S_\phi(u,v)=l_N\circ l_{N-1}\circ\ldots\circ l_1(x(e_i),x(f_j)) \tag{7}$$

In the above calculation formulas, N refers to the layers of the neural network while x(e_i) and x(f_j) means the context-window which is centered by e_i and f_j.

Similar to the common HMM word alignment, as a sentence pair (e,f),we choose Viterbi alignment as the final alignment result.

$$a^*=\arg_a\max p(a|e,f)=\arg_a\max\psi(a|e,f) \tag{8}$$

In order to find Viterbi alignment, we firstly calculate every translation score for each source language and the target word pair, through the neural network.

$$S_{\phi}(u,v) = l_N \circ l_{N-1} \circ \ldots \circ l_1(x(e_i), x(f_j)), i = 0,1,\ldots |e|; j = 1,\ldots,|f_j|; \tag{9}$$

Experiment Results and Analyses

The experimental data contains monolingual data, bilingual parallel corpus and bilingual sentences with manual word alignment. We get all the monolingual data from the Internet. After the process of de-duplication and segmentation, there are about 1.1 billion sentences in English section and 0.3 billion sentences in Chinese section. These monolingual texts are used to train the low-dimensional vector in the model, and used as the character of reference system for the part of Our bilingual parallel corpus contains all bilingual corpus in NIST08 machine translation evaluation and bilingual data mining from the Internet. After the pre-process of de-duplication, bilingual data altogether contains about 260 million bilingual statements. Bilingual parallel corpus is used for training the neural network word alignment model presented from context, and the word alignment model of other reference system. We use the manual word alignment data used in Haghighi[13] , which contains 491 bilingual sentence pairs. Due to differences in Chinese Word segmentation tools, we use some heuristic rules to convert the data into our word segmentation standard. In addition to these 491 words, which are used for the evaluation data, we also have the manual labeling for the 600 words came from FBIS data set (LDC2003E14). This set of 600 words, as the development set, is used to adjust context models and various parameters in the basic system.

Using the method of testing word alignment F-1 component on the annotation data as evaluation criteria. Generally, manual word alignment could be divided into Possible Alignment Link and Sure Alignment Link. For the given bilingual sentence pair, it is assumed that the possible word-alignment connected set for manual labeling is S, and the model prediction for word-alignment connected set is A, so the accuracy, the recall rates and F-1 component of the word-alignment on this bilingual sentence pair will be counting according to the following formula.

$$prec = \frac{|p|}{|A \bigcap P|}, \quad rec = \frac{|S|}{|A \bigcap S|}, \quad f_1 = \frac{prec + rec}{2 \times prec \times rec} \tag{10}$$

The length in the word-aligned neural network is 20. In the formula 3.30, it defines that the network contains two hidden layers, on which the entire network contains superficial layer, three linear layers and two nonlinear layers, and the length of the hidden layers are respectively 120 and 10. In the formula 3.31, it defines that the bi-linear layer network contains a hidden layer, on which the entire network includes superficial layer, a bi-linear layer, a nonlinear layer and a linear layer. And the length of the hidden layer is 4, the length of the context-window for the source language and target language is 3.

Table 1. The results based on the massive experiments of word-alignment.

Setting	Prec.	recall	F-1
HMM	0.768	0.786	0.777
CDDL-HMM	0.810	0.790	0.798

From table 1, the accuracy for our model is higher, and the recalling rate is similar with the results of the reference system. Through the analyses of the results, we found that, according to the reference system and the model presented by this paper, the lack of word alignment is mainly related to the stop

words. Besides, our model improved some low-frequency words, especially on the processing of some proper nouns.

Conclusion and Prospect

This model is trying to combine a multi-layer neural network with an undirected probabilistic graphic, and use the word similarity and the context information to have a more accurate modeling. We also examine the semi-supervised training method and unsupervised training method on the bilingual parallel corpus. After the massive word alignment experiments from English to Chinese, the model which is presented in this paper, compared with the reference system, could significantly improve the effect of word alignment.

The time complexity of the Viterbi deduction for the entire is consistent with the traditional word alignment model. But, in fact, the time for calculating the translation component by the neural network is far more than the time of forward-backward algorithm. So, the time for the deduction used by this model will be longer than the time of traditional word-alignment deduction.

References

[1] Wu D. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Computational linguistics, 1997,23(3):377-103.

[2] Deng Y' Byrne W. HMM word and phrase alignment for statistical machine translation. Audio, Speech, and Language Processing, IEEE Transactions on, 2008.

[3] He X. Using word dependent transition models in HMM based word alignment for statistical machine translation. Proceedings of the Second Workshop on Statistical Machine Translation, 2007.

[4] Liang P, Taskar B, Klein D. Alignment by agreement. Proceedings of NAACL, 2006.

[5] Berger A L, Pietra V J D, Pietra S A D. A maximum entropy approach to natural language processing. Computational Linguistics, 1996.

[6] Moore R C. A discriminative framework for bilingual word alignment. Proceedings of Proc. HLT-EMNLP, 2005.

[7] Koehn P. Statistical machine translation. Cambridge University Press, 2010.

[8]Liu Y. A Shift-Reduce Parsing Algorithm for Phrase-based String-to-Dependency Translation. Proceedings of Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Sofia, Bulgaria: Association for Computational Linguistics, 2013. 1-10.

[9]Hinton Gs Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE, 2012, 29(6):82-97.

[10]Auli M, Galley M,Quirk C, et al. Joint Language and Translation Modeling with Recurrent Neural Networks. Proceedings of Proceedmgs of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA: Association for Computational Linguistics, 2013. 1044-1054.

[11]Devlin J, Zbib R, Huang Z, et al. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. Proceedings of ACL 2014, To appear, 2014.

[12]Feng Y, Cohn T. A Markov Model of Machine Translation using Non-parametric Bayesian Inference. Proceedings of Proc. ACL, 2013.

[13]Haghighi A, Blitzer J, DeNero J, et al. Better word alignments with supervised ITG models. Proceedings of Proc. ACL-IJCNLP, 2009.