Example-based Hybrid Higher-order Neural Network Cognition applied for Archive Translation

CHEN LILAN[1]， CHEN YONGSHENG[2]

1. School of Foreign Languages, Guangdong Pharmaceutical University, Guangzhou, China.
2. School of Information Management, Sun Yat-sen University, Guangzhou 510006, China Corresponding author: Yongsheng Chen.

Abstract. This paper constructs the basic principles and system structure of example-based hybrid higher-order neural network cognition for machine translation. On this basis, the paper elaborates the construction method of English and Chinese bilingual archive instance library, and gives the multi-granularity alignment techniques (paragraph level, sentence level and word level) used in the processing of bilingual archive corpus, forming a multi-level alignment method and improving the accuracy of bilingual archive corpus alignment. Then, the paper discusses and gives the algorithm of sentence similarity based on words in archive databases and the retrieval algorithm of the most similar instances. Finally, the paper makes an experiment on the example-based machine translation (EBMT) of archives, evaluates the performance of the translation, analyzes the results, concluding the advantages and disadvantages of this method. The experiment results show that the expected target of archive translation is achieved, that is, the simple archive translation can be performed, and when there are sentences in the instance library that are similar to the sentences to be translated, the archive translation of higher quality can be obtained.

Keywords: archive, hybrid higher-order neural network, cognition, bilingual archive instance library, machine translation of archives

1 Introduction

It accelerates and expands the spread of information around the world, being increasingly important in today's economic globalization and showing self-evident research significance. Machine translation automatically converts one natural language

into another with the use of computers [1]. The research on machine translation can be traced back to the birth of the first electronic computer. Ever since then, people have started to pay extensive attention to machine translation in the research field of natural language information processing [2]. Translation is in essence an intelligent activity, which, from receiving the source language, first analyzes the source language, then converts the source language into the target language according to complex inference rules, and finally generates texts that conform to the grammar of the target language, that is, the final translation. The series of complex behaviors and reasoning in the process of translation requires a large and broad knowledge base (including language-related and language-irrelevant background knowledge). Also, machines are expected to have the function of knowledge base deep learning and applying it to reasoning. The research of machine translation involves many disciplines such as computer science, linguistics, mathematics [3]. It is said that if artificial intelligence problems cannot be absolutely broken through, it is impossible for machine translation to achieve perfect results. It can also be said that one of the ultimate goals of artificial intelligence is machine translation [4-5]. From this perspective, it is self-evident that the study of machine translation is of great value and challenges[6-8].

In this paper, in accordance with the characteristics of archives, the first part introduces the progress of archive translation, and the second part constructs the basic principles and system structure of example-based hybrid higher-order neural network cognition. On the basis of this, the paper proposes a method to construct the instance library of bilingual archives in English and Chinese, and gives the alignment techniques of different granularity (paragraph level, sentence level and word level) used in the processing of bilingual corpus of archives, which forms a multi-level alignment method and improves the accuracy of bilingual archive corpus alignment. The third part discusses and provides the hybrid higher-order neural network cognition algorithm of sentence similarity based on words the most similar instance retrieval algorithm. Finally, the paper conducts experiments on the example-based hybrid higher-order neural network machine translation [10,11] of archives, evaluates the translation performance, and analyzes the experiment results, summarizing the advantages and disadvantages of the method. The experimental results show that the expected target of archive translation is achieved, that is, the simple archive translation can be made, and when there are sentences in the instance library that are similar to the sentences to be translated, the archive translation of higher quality can be obtained.

2 . Example-based Hybrid Higher-order Neural Network Cognition

2.1 The Translation Structure of Examples

We can get some inspiration from the human translation, and the EBMT is to imitate this process. EBMT also belongs to corpus-based machine translation, that is to say, align the instances in large-scale parallel corpus to establish translation instance

library, which is made up of two fields, one storing the source sentences, the other storing the parallel translated texts, and on this basis, an algorithm for existing sentence searching and similar sentence matching is established. This is the general process of instance-based machine translation [12-13].

Compared with other machine translation methods, EBMT method has more advantages than other machine translation methods, which are mainly listed in the following aspects:

①It's more convenient for the system maintenance. Compared with the traditional rule-based machine translation, the tedious and large knowledge base is replaced by a large number of parallel contrast example sentence pairs, and it is easy to add and delete the items in the instance library.

②The translation is of high quality, particularly when it comes to the input with high matching degree in the instance library, the translation is highly accurate.

③It reduces the relevance of language knowledge. The EBMT method actually uses similarity algorithm to match sentence pairs in the instance library without linguistic analysis of the language in the instance library. By summing up the experience and lessons from various practices, we draw the conclusion that in the study of machine translation of archives, bilingual languages belong to two different language families with very different grammar.

Based on the above analysis, it is safe to say that the EBMT for archive translation is comparatively practical and effective.

2.2 The Basic Principles

As mentioned in the previous section, the EBMT approach essentially imitates the process of human translation. From the perspective of specifications, the principle of EBMT is as follows. When the system receives a source sentence, it first divides the sentence into phrase fragments in accordance with the set rules, and then compares these fragments with contrast examples in the instance library, finding out the most similar matching fragments in source language, and finally reorganizes these fragments into sentences, also our target translated texts [14]. In brief, EBMT can be roughly divided into three steps: matching, aligning and reorganizing, respectively corresponding to the three steps of analysis, transition and generation in traditional rule-based machine translation [15,16,17]. The following is an example to illustrate the principle. Suppose there are two instances A and B in the instance library as follows:

A. Canton Events and Current Rumors　　各项时事传闻录

B. Destruction confirmation of the archival department　档案形成部门同意销毁确认

Now type in the source expressions needing translation: Canton, Events, Current, Rumors; destruction, confirmation, archive, department.

The translation process of the EBMT system is as follows:

(1) the matching stage, in which the expressions “destruction, confirmation, archives, department” in the sentence to be translated successfully find their matches in the source sentence in the instance sentence pair B, and the expressions “Canton, Events, Current, Rumors” in the sentence to be translated can find their matches in the source sentence in instance sentence pair A.

(2) the aligning stage, in which it can be seen from the matching information in the instance library that translation fragment corresponding to the matching part in the instance sentence pair A is “and”, and the translation fragment corresponding to the matching part in the instance sentence pair B is “forming department”.

(3) the reorganizing stage, in which the translation of the two fragments obtained in stage (2) is combined into the final translation.

Through the illustration of the above example, we can clearly see the EBMT method is comparatively advantageous over the rule-based machine translation method, in addition, since instances have been stored in the parallel corpus, it is not necessary to retranslate sentences with the same or highly similar example sentences in the instance library [18,19].

2.2　Construction of the Hybrid Higher-order Neural Network System (HHONNS)

The hybrid higher-order neural network refers to the network of many different types of higher-order interconnected neurons, in which the power parameter in each neuron calculation formula is different, i.e., neurons are in the multidimensional space with different geometric shapes. As for the problem of diverse example space and bilingual corpus shapes in EBMT, the HHONNS can be applied to deal with the problem, with the advantage that the number of neurons used may be reduced after intuitive training and the disadvantage of long learning time. In the following part the

retrieval training algorithm of instances based on this network is introduced in detail[20,21,22].

In step 1, the example space bilingual corpus is divided into A word bank. First, the center vector of a certain class of higher-order neurons can be selected in a variety of ways, and the neuron is labeled as A word bank.

In step 2, example words in this bank are sent to the neuron one by one through the input node for calculation, and sort the calculated results in descending order, and the result is denoted as Oi, i = 0, 1... n - 1

In step 3, find the heterogeneous word examples with the smallest Euclidean distance from the center vector in the descending order from large to small

Suppose that its position in Oi is d, a point between Od and Od-1 is selected as the threshold value of the current neuron, whose position can be determined by the formula $\theta = \alpha \mathrm{O}_{\mathrm{d}} + (1-d)O_{d-1}$, and the number and threshold value of the example word banks divided by the neuron are recorded.

In step 4, select other types of higher-order neurons as candidate neurons and repeat Step 2 and Step 3. When all neurons are calculated, the calculation results are recorded, and select the type of neurons which divide the largest number of the same type of work bank examples as the network neuron. Of course, this method may not be optimal.

In step 5, delete the word bank example divided by the current network neuron from the training example set and start to adjust the direction of the neuron.

In step 6, initialize each connection weight of W to be 1 with the current neuron as the core. Adjust the weight of W to the direction smaller than 1. In other words, neurons extend towards this dimension to include more samples of the same kind, and the modification of weights can be calculated according to the formula $\mathrm{W}_t(t) = \mathrm{W}_t(t-1) - \eta.\lambda$.

In step 7, calculate the distance between other word examples and the word example after adding W weight vector, judge whether any new word examples are covered, and if so, judge whether the new example is similar to the core example. If it is in the same category as the core example, record the W and continue to adjust; if it is not in the same category, return to the original W and continue to adjust in other dimensions.

Step 8, when W is adjusted to a certain value or the number of adjustments reaches a certain figure, solidify the neuron, i.e., the neuron is no longer extended. Record the newly divided word bank samples in accordance with W and delete them.

Step 9, take the remaining word banks as the new sample set. Repeat Step 2 to Step 8 until there is only one type of examples in the word bank.

Step 10, construct a higher-order neuron model to divide the remaining word examples.

Step 11, number all higher-order neurons in sequence. The neuron learned first has a small serial number and the neuron learned later has a large serial number. Suppose the number of higher-order neurons is H, then the serial number is 1 ~ H, and the transfer function of the higher-order neuron is changed to a hard-limiting function.

3 The Experiment and Analysis of the Experiment Results

The Experiment Process:This paper studies the EBMT of Canton Maritime Customs archives stored in Guangdong Provincial Archives, which is a corpus-based neural network machine translation method. Therefore, before the translation, it is necessary to preprocess the corpus to form the corresponding instance library. The processing goes as follows: (1) 30,000 sentences in Canton Maritime Customs archives are obtained through the Guangdong Provincial Archives corpus, with subjects

The Translation Process:The most critical instance library in the experiment is established through the corpus preprocessing module in the previous step, and the next step is to translate. This module is written in C++ language, aiming at completing the preliminary machine translation of Canton Maritime Customs archives in Guangdong Provincial Archives between Chinese and English.

The experiment is completed by the combination of the inverted index, computing sentence similarity, retrieving the most similar sentences.

The specific process of translation is as follows:

(1) Processing of Source Statements

Type in the current source sentences that need to be translated, in this case, the archival sentences in English. Sentences are segmented, providing the foundation for the subsequent neural network similarity calculation of these sentences.

Retrieval of Similar Instance Sentences

Search for similar sentences from the instance library. The sentences obtained after word segmentation are calculated their similarities and sorted according to their calculation results by using the word-based sentence similarity training formula based on the constructed hybrid higher-order neural network. Here, we set a threshold value of 0.8, taking sentences with similarity over 0.8 as the same as the sentences to be translated, and directly putting out the target language corresponding to the similar sentences as the translation result. When there are not very similar sentences, but there are locally similar sentences (that is, sentence similarity is less than the threshold), we proceed to the next step.

(3) Reorganization of Instance Fragments.

The Translation Results: Table 3-1 lists some examples of Canton Maritime Customs archive translation in Guangdong Provincial Archives in this experiment

Input （English）	OUTPUT(Chinese)
When I assumed the post of Inspector General of Customs one of the cardinal features ofmy policy was the introduction of the principle of equality of opportunity for qualified Chinese …	本总税务司自就职以来，以给与合格华员平等机会，为政策主要特点之一厄
I have now to circulate, for your information and guidance, copy of Kuan-wu Shu despatch No. 15763, from which you will see that Mr. Loy Chang (郑莱) has been appointed Acting Director General of the Kuan-wu Shu.	兹奉关务署第 15763 号训令内开，委派郑莱先生为财政部代理关务署署长。仰各关知照。
Finally, you are informed that the Government have recently conferred upon the Inspector General the Order of the Brilliant Jade with Blue Sash (3rd Class) …	最后，政府近授总税务司三等蓝色大绶采玉勋章厄
… it is necessary to notify the Service that the number of applications for voluntary retirement has recently increased to abnormal proportions, …	……须通知海关，近来申请自愿退休人数增加到不正常的比例……
The Civil Governor Li Yao-han has gone to Hongkong from Shiuhing for a short stay there.	省长李耀汉已经从肇庆去了香港，并要在那里作短暂逗留。
The new Acting Civil Governor Chai Wang has issued a notification informing the public that he will take over the seals of office at noon tomorrow, the 20th instant.	新任代理省长翟汪已发出通告周知民众，明天(10 月 20 日)中午他将接印上任。
T'ang Chi-yao has left Kobe on the 15th instant for China.	唐继尧已于本月 15 日离开神户回国。
Pei Tsu-i, Sub-Manager of the Bank of China, Canton, returned today from Hongkong.	广州中国银行副经理贝祖贻今天从香港回穗。
The Central Government has issued a Mandate strictly forbidding the smuggling of opium by military officers.	中央政府已公布训令，严禁军官进行鸦片走私。
It is said that the Allied will present a second note to China on 2nd January 1919.	据说，协约国准备于 1919 年 1 月 2 日再次向中国提出照会。

4 Conclusion

This paper starts from basic work such as the construction of instance library and word alignment, finally verifies the feasibility of the example-based hybrid higher-order

neural network machine translation method in bilingual archive translation through the sentence similarity calculation of hybrid higher-order neural network and the combination of similar sentences. Finally, from the analysis of the experiment results, the “shining point” in using example-based hybrid higher-order neural network machine translation method in Guangdong Customs archive translation is very obvious. By using this method, the tedious grammatical analysis of English sentences and Chinese sentences is abandoned, which can bring convenience for translation between any languages, as high-quality translation can be obtained just by storing the preprocessed bilingual sentences in the instance library and then retrieving similar instances.

Acknowledgments. This work was supported by one of the Major Projects of National Social Science Fund of China No. 17ZDA200.

References

1. Wang, R., Zhang, W., Shi, Y., Wang, X., & Cao, W. (2019). GA-ORB: A new efficient feature extraction algorithm for multispectral images based on geometric algebra. IEEE access, 7, 71235-71244..
2. Hutchins, W. J., & Somers, H. L. (1992). An introduction to machine translation (Vol. 362). London: Academic Press.
3. Gu, J., Neubig, G., Cho, K., & Li, V. O. (2016). Learning to translate in real-time with neural machine translation. arXiv preprint arXiv:1610.00388.
4. Vilar, David, Jan-T. Peter, and Hermann Ney. "Can we translate letters?." Proceedings of the Second Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2007.
5. Cao, Wenming, et al. "Content-based image retrieval using high-dimensional information geometry." Science China Information Sciences 57.7 (2014): 1-11.
6. Brown P F, Cocke J, Della Pietra S A, et al. A statistical approach to machine translation[J]. Computational linguistics, 1990, 16(2): 79-85.
7. Och, Franz Josef. "Minimum error rate training in statistical machine translation." Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003.
8. Chang P C, Galley M, Manning C D. Optimizing Chinese word segmentation for machine translation performance[C]//Proceedings of the third workshop on statistical machine translation. Association for Computational Linguistics, 2008: 224-232.

9.Cao, Wenming, Feng Hao, and Shoujue Wang. "The application of DBF neural networks for object recognition." Information Sciences 160.1-4 (2004): 153-160.

10.Cao W, Pan X, Wang S. Continuous speech research based on two-weight neural network[C]//International Symposium on Neural Networks. Springer, Berlin, Heidelberg, 2005: 345-350.

11.Wang S. A new development on ANN in China—biomimetic pattern recognition and multi weight vector neurons[C]//International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing. Springer, Berlin, Heidelberg, 2003: 35-43.

12. Cao W M, Feng H, Zhang D M, et al. An adaptive controller for a class of nonlinear system using direction basis function[C]//6th International Conference on Signal Processing, 2002. IEEE, 2002, 1: 54-57..

13.Ramon J. Clustering and instance based learning in first order logic[J]. AI Communications, 2002, 15(4): 217-218.

14.Tirkkonen-Condit S. Unique items-over-or under-represented in translated language?[J]. Benjamins Translation Library, 2004, 48: 177-186.

15.Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., ... & Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. Machine translation, 25(2), 127-144.

16.Simard M, Ueffing N, Isabelle P, et al. Rule-based translation with statistical phrase-based post-editing[C]//Proceedings of the Second Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2007: 203-206.

17. Wang, R., He, Y., Huang, C., Wang, X., & Cao, W. (2019). A novel least-mean kurtosis adaptive filtering algorithm based on geometric algebra. IEEE access, 7, 78298-78310..

18.Baudette M, Castro M, Rabuzin T, et al. OpenIPSL: Open-Instance Power System Library—Update 1.5 to "iTesla Power Systems Library (iPSL): A Modelica library for phasor time-domain simulations"[J]. SoftwareX, 2018, 7: 34-36.

19.Derrac J, García S, Herrera F. A survey on evolutionary instance selection and generation[M]//Modeling, Analysis, and Applications in Metaheuristic Computing: Advancements and Trends. IGI Global, 2012: 233-266.

20. Shen M, Wang R, Cao W. Joint sparse representation model for multi-channel image based on reduced geometric algebra[J]. IEEE access, 2018, 6: 24213-24223..

21.Wang, Rui, Yijie Shi, and Wenming Cao. "GA-SURF: A new Speeded-Up robust feature extraction algorithm for multispectral images based on geometric algebra." Pattern Recognition Letters 127 (2019): 11-17.