Research on Tibetan Question Answering Technology Based on Deep Learning

Shengyin Zhu\textsuperscript{1,a} Xiangzhen He\textsuperscript{2,b,*} Bingjun La\textsuperscript{3,c} Hongzhi Yu\textsuperscript{4,d}

\textsuperscript{1}Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education
Northwest Minzu University
Lanzhou, Gansu
730000, China

\textsuperscript{2}Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education Northwest Minzu University
Lanzhou, Gansu
730000, China

\textsuperscript{3}Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education Northwest Minzu University
Lanzhou, Gansu
730000, China

\textsuperscript{4}Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education Northwest Minzu University
Lanzhou, Gansu
730000, China

\textsuperscript{a}87961192@qq.com \textsuperscript{b}hexiangzhen@163.com \textsuperscript{c}17154909@qq.com \textsuperscript{d}yuongzhi@hotmail.com

ABSTRACT: By applying the theory, technology and system of the Chinese Question Answering system, combining with the background of large data and cloud processing, integrating the theoretical knowledge of Tibetan linguistics, and covering computer high-tech of Tibetan machine translation, neural network learning, problem analysis, classification and answer generation, etc, this paper explores the optimization model of the new Tibetan-Chinese bilingual Question Answering system. Finally, the Tibetan/Chinese bilingual Question Answering system based on deep learning is completed to provide technical support for the economic development of ethnic minority areas.

1. INTRODUCTION
In the era of big data and cloud computing, it is an important trend of the development of national information technology to develop various network information services around the mobile Internet. In ethnic areas, the sharing of information resources in the Tibetan language is a major demand for national development. With the rapid development and popularization of the Internet as well as the
change of the way of acquiring information resources, the daily information that users come into contact with has exploded. However, the traditional information processing method can’t meet the practical needs of the masses in their daily life for the processing, classification and retrieval functions of various kinds of Tibetan language information.

How to find relevant content quickly and effectively in the increasingly expanding information has become a hot issue. At the same time, deep neural networks have been deeply explored and achieved remarkable results in tasks such as image classification and speech recognition, showing excellent learning ability. Therefore, in the face of the disadvantages of traditional information retrieval, in order to meet users’ high-level search needs, the information retrieval method with deep learning ability has gradually become a new research trend.

2. RESEARCH STATUS AT HOME AND ABROAD
In recent years, the remarkable development of deep learning and natural language processing technology has made it possible to build more complex applications. The Question Answering system has attracted a lot of attention recently. According to the fields involved, the Question Answering system can be divided into two types: open fields and task-oriented.

2.1 Deep Learning
In 2011, Microsoft speech recognition uses deep learning technology to reduce the error rate of the speech recognition by 20-30%, which is the biggest breakthrough in the field for more than a decade. In 2012, Google's Google Brain project built a self-learning neural network with more than 1 billion nodes using a server cluster of 16,000 processors, which automatically extracted the concept system from the large amount of input data. Chinese search giant Baidu has also set up a deep learning lab.

2.2 Automatic Questions and Answers
The public information service platform, faced with the processing and feedback of users’ input information, is realized through the computer automatic reasoning and answer extraction technology within the system, namely the automatic Question Answering system. The international Question Answering system is mainly for English, among which the more mature Question Answering system includes START, Answer Bus, Ask Jeeves, Watson, etc.

Domestically, there are mainly institutes of computing of the Chinese academy of sciences, Fudan University, Harbin Institute of Technology, Peking University, Beijing Language and Culture University, Taiwan University, Taiwan Academia Sinica and other units engaged in the study of computer Question Answering system. NKI (National Knowledge Infrastructure) question-and-answer system was developed by the Institute of Computing of the Chinese Academy of Sciences. Users can query natural language methods on National Geographic Knowledge Base, Weather Forecast Knowledge Base, People Knowledge Base, TCM Disease Knowledge Base and other knowledge base through natural language questioning.

2.3 Machine Translation
On the information public service platform, both the intelligent computer response and the realization of its public opinion cloud function are closely connected with the machine translation technology. Machine translation is to use computers to realize the translation between different languages, which is the process of conversion from the source language to the target language.

Most of the research on machine translation involving Tibetan language focuses on phrase-based machine translation. At present, there are some research units in China, such as Xinjiang University, Northwest Minzu University, Minzu University of China, Institute of Computing Technology of Chinese Academy of Sciences and Institute of Automation of Chinese Academy of Sciences.

To sum up, the Question Answering system based on deep learning is mostly used in Chinese or English, mainly in Tibetan, and the bilingual Question Answering technology supplemented by Chinese has not been applied yet, and is still under exploration.
3. TIBETAN AUTOMATIC QUESTION ANSWERING SYSTEM BASED ON DEEP LEARNING

Through the research of deep neural network-based semantic representation, Tibetan/Chinese translation, language problem solving and answer generation, the Question Answering system can more accurately understand user's questions in natural language form and returns a concise, precise matching answer by retrieving the heterogeneous corpus or the question answering knowledge base. Compared with the search engine, the Question Answering system can better understand the true intention of the user's question, and more effectively meet the user's information needs.

Figure 1. Q & A system processing framework

3.1 Question Interpretation

Question comprehension is a key part of the Question Answering system to understand the user's intention. The performance of the question understanding module directly restricts the effect of the subsequent processing module. Question understanding mainly includes question classification, topic focus extraction, and problem extension processing.

3.1.1 Language-oriented multi-information fusion for Tibetan/Chinese translation

The Tibetan/Chinese machine translation is moving from the traditional phrase-based translation to the syntactic translation. Based on Tibetan phrase syntax, this project integrates semantic cues and conducts hierarchical and systematic research on the Tibetan/Chinese machine translation. By using the two major clues of Tibetan syntax and semantics, the knowledge rules are integrated into the phrase syntax tree to provide support for Tibetan/Chinese machine translation and realize Tibetan/Chinese machine translation at a higher level than lexical-level and phrase-level, as shown in figure 2.
3.1.2 Semantic representation method based on deep neural network

3.1.2.1 Semantic representation method based on convolutional neural network (CNN)
The semantic representation learning based on CNN is to scan the sentence through CNN, extract features, select features, and finally combine them into the representation vector of the sentence. First, the sentence is scanned from left to right with a sliding window. Each sliding window has multiple words, and each word is represented by a vector. Feature extraction is performed by a convolution operation in the sliding window. In this way, a series of features are obtained at various locations. Then, the feature is selected by the max pooling operation. Repeat the above operation many times to get multiple vector representations, and connect these vectors to get the semantic representation of the whole sentence.

3.1.2.2 Semantic representation based on RNN
Sentence modeling based on RNN refers to a sentence as a sequence of words. Each word is represented by a vector, and each position has a middle representation, which consists of vectors, representing the semantics from the beginning of the sentence to this position. It is assumed that the intermediate representation of each position is determined by the word vector of the current position and the middle representation of the previous position, and modeled by a cyclic neural network. RNN treats the middle representation at the end of a sentence as the semantic representation of the whole sentence.

3.2 Information Retrieval
According to the query expression obtained by the question understanding, the information retrieval module is responsible for retrieving related information from the heterogeneous corpus and the question answering knowledge base, and transmitting the relevant information to the subsequent answer generation processing module. For different Question Answering systems, the retrieval model and the retrieval data form of the system are also different. For the Question Answering system based
on free text data, the information retrieval process is a filtering process that gradually narrows the answer range, mainly including document retrieval and paragraph sentence group retrieval. For the Question Answering system based on the question answer pair, the information retrieval process is to obtain a candidate question similar to the user's question by querying the question, and return the corresponding candidate answer list.

3.2.1 Tibetan keywords extraction and expansion
Stop words appear in a large number of texts. Syntactic analysis is carried out for Tibetan language, and stop word list is used to filter words that have no relation to problem analysis. Keywords are extracted by combining word clustering algorithm of documents, implicit subject model and document structure information. The unsupervised similarity calculation method is adopted to calculate the semantic distance of the key words and extend it to answer the questions.

3.2.2 Shallow Semantic Analysis of Tibetan Language
Based on the phrase structure role labeling system, according to the heuristic rules, the deleted part cannot be the syntactic component of the semantic role, and judge whether a labeling unit is a semantic role; the semantic role analysis is regarded as a classification process, and the adopted features include local information such as predicate, path, phrase type, position, state, voice, core word, as well as global features such as marking unit semantic role sequence features and number of target verbs. The machine learning models such as conditional random field, maximum entropy and support vector machine are used for feature learning, classification and semantic analysis.

3.3 Problem Reasoning and Answer Generation Based on Multi-source Knowledge Association

3.3.1 Analysis of question
The classification of Tibetan question sentences begins with modal particle filtering, word segmentation and the recognition of named entity, extracts query words and identifies question categories, and uses the comparison table and synonym dictionary as features to perform query expansion and generate query vectors.

The problem classification is treated as a mapping process, and the question of the unlabeled category is mapped to the defined semantic category according to its answer type. The mapping can be a one-to-one mapping or a one-to-many mapping. For questions with question words, we can use rules to classify questions by creating a list of questions. For questions that do not contain question words, according to different question types, modeling and recognition based on the classification learning model is adopted.

3.3.2 Complex problem decomposition
The Tibetan question input is based on grammar analysis and statistical classification method to determine whether the problem is decomposed. The domain knowledge base is used as the knowledge source to classify the problem, and to determine how to decompose is easier to get the answer. After the problem is decomposed, sub-problem sets are formed, and the final alternative answer is synthesized through answer synthesis. The decomposition process is as follows: structure the problem, refine and decompose the problem, and accurately describe the problems, divide the complex problem of the bad structure into a good structural problem, and use the existing structural method to solve the problem to the maximum extent; Representation, organization, and reasoning of domain knowledge, refine the problem, clarify the composition of the problem and the relationship between them, and establish a database, model library, method library, knowledge base, case library, text library, problem library, and the links between them; With the help of the formalized problem representation and the problem reduction theory of the knowledge representation method in artificial intelligence, the complex decision-making problem is reduced to the set of sub-problems, and the relatively abstract
decision-making problem of a whole is manifested into several relatively independent sub-modules, thus the relatively complex problem is decomposed into relatively simple sub-problems.

Figure 4. Decomposition of complex problems

3.4 Answer Extraction Based on Tibetan FAQ

In order to improve the processing performance of the answer to the question, the FAQ library of the Tibetan language is established for the common questions input by the user. If the question input by the user can be matched with the question-answer pair in the FAQ library, the corresponding answer in the FAQ library is directly output.

Figure 5. The answer extraction process based on the Tibetan FAQ library

3.4.1 Tibetan FAQ construction

According to the theory of linguistics, people use interrogative sentences to inquire information. Interrogative sentences are generally divided into yes or no questions, choice questions, and specific questions. In the research of question processing, the emphasis is usually put on the special interrogative sentences. Therefore, this subject chooses the special questions as the main research focus.

When establishing the Tibetan language FAQ library, it is generally built according to the question-answer pair. According to different languages, manual translation, collection and the automatic construction of knowledge base, all question-answer pairs are divided into six categories according to questions, including time, entity, task place, quantity and description, and corresponding labeling marks are studied.

According to the user's satisfactory answer to the question, the question is stored in the Tibetan FAQ library, and the content is continuously updated.

3.4.2 Extract answers from the Tibetan FAQ

The reason why we deal with the frequently asked questions separately is that the answers in the question bank are fixed. Therefore, for the questions raised by users, we only need to find similar questions in the FAQ library. The method adopted is to calculate the similarity of questions. If similar sentences can be found, the answer is returned directly, which can greatly improve the efficiency of the system and the accuracy of the return answer. Because the user's question has many different forms, it is necessary to extend the semantic of the key words of the question asked by the user before
calculating the similarity of the question, and then we study the corresponding similarity algorithm to calculate it. If similar problem cannot be found, the problem is transferred to complex problem solving.

3.5 AnswerFinal Fusion and Sorting Strategy
For the fusion of Tibetan and Chinese bilinguals, there are Tibetan and Chinese answers in the alternative answers, so the parameters of Tibetan and Chinese bilingual answers need to be adjusted and weighted with certain parameters. With the help of multi-source knowledge base, the normalization is finally carried out to make the score of the two answers more reasonable.

Different alternative answers to a question may take many forms, but they all mean the same thing. Ranking techniques are very confusing to exploit the relative differences between alternative answers. If there is no merging, the rank algorithm wastes time unnecessarily trying to distinguish answers that have the same meaning but different form. After combining the matching, standardization and co-definition digestion algorithms, the equivalent or related candidate answers can be identified, such as "Monday" and "Monday", and the corresponding features can be combined to obtain the final score of the candidate answers. Among them, "co-fingering" is the action of merging different descriptions of the same entity in the article.

4. CONCLUSION
By applying the theory, technology and system of the Chinese Question Answering system, combining with the background of large data and cloud processing, integrating the theoretical knowledge of Tibetan linguistics, and covering computer high-tech of Tibetan machine translation, neural network learning, problem analysis, classification and answer generation, the optimization model of the new Tibetan-Chinese bilingual Question Answering system is explored and the Tibetan/Chinese bilingual Question Answering system based on deep learning is completed to provide technical support for the economic development of ethnic minority areas.

Under the background of “One Road, One Belt”, it is urgent to study Tibetan/Chinese bilingual question answering technology based on deep learning, which is beneficial to eliminate the disadvantages of national areas in terms of stability and rapid development, and is conducive to the realization of the “multi-integration” architecture in the field of Tibetan language information services in the digital era. The study of Tibetan/Chinese bilingual question answering technology provide some basic theory on the strategic arrangements concerning the management of digital information in minority Tibetan languages, this fundamentally safeguard the national unity of the country, social stability and the construction of a harmonious society.

ACKNOWLEDGEMENTS
This work was supported by Gansu Province Education Science Planning Project (No.GS[2016]GHB0203).

REFERENCES
[1] Douglas W.Oard. Cross-language Information Retrieval. Annual Review of Information Science and Technology. 1998(6).221-230
[2] Hou Yanfei. Research on Cross-Language Information Retrieval[D]. Peking University Dissertation, 2003.
[3] Cao Shujin. Foreign Information Retrieval Language and Natural Language Retrieval[J]. Journal of Information Science (Volume 3).1998.134-135.
[4] Language and machines: computers in translation and linguistics (ALPAC report). National academy of sciences, Nat'l Academies Press. 1966.329-344. [5] W. Hutchins. Machine Translation: Past, Present, Future. Ellis-Horwood Limited, Chichester, England. 1986.34-36
[5] Jianfeng Gao, Jian-Yun Nie. TREC-9 CLIR Experiments at MSRCN.234-237
[6] Zhang Yuejie et al. Cross-language information retrieval based on machine translation [J]. Small microcomputer system.2004(7).23-25
[7] S. Yu. Automatic evaluation of quality for machine translation systems. Machine translation. 1993, 8: 117-126
[8] Wang Jin et al. Cross language information retrieval model based on Ontology [J]. Journal of Chinese Information Processing. 2004(3).35-36
[9] Luo Weihua. Overview of Cross-Language Information Retrieval Methods[J]. Digital Library Forum, 2006 (9).1-7.
[10] Ren Chengmei. Development and Prospect of Cross-Language Information Retrieval[J]. Library Science Research, 2006 (4).79–82.
[11] Ballesteros, L. and Croft, B. Dictionary Methods for Cross-Lingual Information Retrieval. Proceedings of the 7th International DEXA Conference on Database and Expert Systems. pages 791-801.1996.
[12] Fung, P. and K. McKeown. 1997a: A technical word- and term-translation aid using noisy parallel corpora across language groups. Machine Translation 12,53–87.
[13] Wang hao. Discussion on the Implementation Method and Key Technologies of Cross-Language Information Retrieval[J]. Journal of Information, 2005(7): 46–49.
[14] Bonnie J. Dorr. Interlingual machine translation A parameterized approach. Artificial Intelligence Volume 63, Issues 1-2, October 1993, Pages 429-492.