CHINESE NAMED ENTITY RECOGNITION BASED ON KNOWLEDGE-BASED QUESTION

Arvind Chandrasekaran

Colorado Technical University, USA

ABSTRACT

The knowledge-Based Question Answering (KBQA) system is an essential part of the customer service system aiming to answer natural language questions by recovering the structured data stored under the knowledge base. KBQA answers the natural language questions by recovering the structured data stored under the knowledge base. KBQA receives the user's query and first needs to recognize the topic for the query entities like the location, name, organization, etc., The process is Named Entity Recognition (NER) using the Bidirectional Long Short-Term Memory Conditional Random Field model, and the SoftLexicon method is introduced as the Chinese NER tasks. A fuzzy matching module is proposed to analyze the application scenario characteristics using multiple methods. The module efficiently modifies the error recognition results, improving entity recognition performance. The fuzzy matching and the NER model are combined into the NER system. The power grid-related original data is collected to improvise the system following the power grid data characteristics.

KEYWORDS

Knowledge-Based Question Answering; SoftLexicon method; Knowledge graph learning representation; System OfQuestion answering; Knowledge; Spatial Temporal;

1. INTRODUCTION

Traditional NER methods depend on the additional features, and it is difficult to mine the semantic information, leading to poor NER model performance when the Out of Vocabulary problem occurs. Deep Learning NER methods have gradually become established through continuous deep learning technology.

Deep Learning NER learns the semantic features through complex non-linear transformation. (Cheng, Pan, Qiao. et al., 2022) The technology's large-scale application has improved the efficiency and accuracy of NER. The Continuous Bag of Words is proposed, and the Skip grams model leads to adoptingthe word-embedding methods and widespread adoption of dealing with Natural Language Processing tasks. Chinese NER is challenging to infer as Chinese sentences can't be segmented naturally, like English by space. For Chinese NER, the common practice would be using the Chinese Word Segment tool for word segmentation before applying the word sequence label. (Cheng, Pan, Qiao. et al., 2022) The errors significantly impact the NER model's performance. The researcher proposes the effective method as the Long Short-Term Memory (LSTM) structure using the word and character information for the Chinese text. The model improvises using the Recurrent Neural Network for the Lattice LSTM. The model focuses on the Chinese NER task implementation for the KBQA scene.

David C. Wyld et al. (Eds): SIPR, NCWC, BIBC, SOFEA, DSML, ARIA, NLP, CSEN -2023
pp. 131-135, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131910

2. PREPARATION OF KNOWLEDGE BASE

Freebase is the most extensive semantic knowledge base; every knowledge piece is the triple entity relation entity. The part of FB5M, Freebase, and more than 7.5k relationships extracts the user knowledge. (Guo, Liu, Wang. et al., May 2020) Considering human memory limitation and the experimentation costs, the knowledge from the Freebase is removed to form the raw materials. Breadth- first search or Depth-first search would make the knowledge graph centralized. The random walk search algorithm is used to avoid the above problems. (Guo, Liu, Wang. et al., May 2020) The extraction strategy is the node selection under the knowledge graph, then recording the node relationship randomly. Then it is switched to the adjacent node with definite likelihood for the random choice or stays in the original node for picking another relationship.

3. DIALOGUE MODEL

Building the agents and design is the procedure of questioning and answering. Under the model, every agent needs to answer the question of others and then decide to ask the following question or end the conversation. (Guo, Liu, Wang. et al., May 2020) The dialogue starts with one agent asking about the selected node and ends when one of the agents can't find the answer or decides not to continue. Every agent comprises two Deep Q-Network (DQN) networks, one to answer and the other to question. DQN network includes fully connected layers. The first layer's dimension corresponds with the length of the current dialogue's feature vector. (Guo, Liu, Wang. et al., May 2020) The network output decides the knowledge base for answering the question and the other on which the current node's relation is used for asking the next question. Agents get trained by questioning and answering. The pictorial representation of the model structure is displayed below.

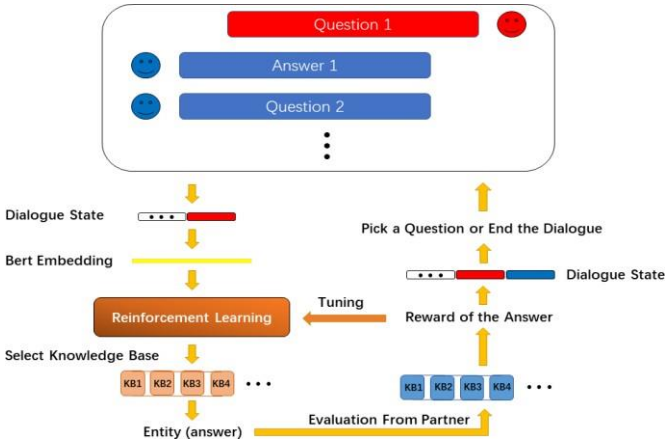

Figure 1. Model Structure

4. CONSTRUCTION OF HETEROGENOUS SUBGRAPH

Nodes are connected based on the relation, subject, and object sequence. The relations are regarded as graph nodes. Further, the connections are explored, starting with the subject, relation, and later object for other nodes, like the relation-based, subject-based, fully connected, and object-based. (Chen, Huang; Yen.et al., October 2022) The path to reach the desired nodes might be shorter with the given different subgraphs. For instance, Narita Airport requires one hop under the S-based compared with two hops under the C-based subgraph. (Chen, Huang; Yen. et al.,

October 2022) Under the F-connected subgraph, every node would need one hop to reach the other nodes.

5. TRANSE-QA ALGORITHM IDEAS

The knowledge base comprises facts played as triples where the object and subject are the entities, and the relationship describes the link type between the object and subject. When the question is mapped with the single triple by the question-answering system, the type of question-answering system is termed Simple QA (the typical Knowledge Base Question Answering (KBQA)). (Chen, He; Wang. et al., May 2020) Every fact or question could be expressed formally as a relationship, head entity, or tail entity. The simple QA question involves a single relationship; for instance., The question of What is Obama's hometown? It can be rephrased as "Obama, birthplace, Honolulu," while the relationship is the birthplace. The question is the relation between the triple and the head entity, and the answer is formed through the tail entity.

Hence the question is transformed into (Obama, birthplace,?), where? represents the answer to the question. (Chen, He; Wang. et al., May 2020) Fetching a single fact from the knowledge base is requiredto answer the question correctly. Simple QA involves simplifying the reasoning process, bypassing the complex reasoning, and clarifying the structure and algorithm. But still, the difficulty in the fact selection is faced, requiring the question to be matched correctly with the knowledge graph facts. (Chen, He; Wang. et al., May 2020) In this regard, the TransE-QA algorithm is an effective tool for solving the aboveproblem. The pictorial diagram for the TransE-QA algorithm is shown below

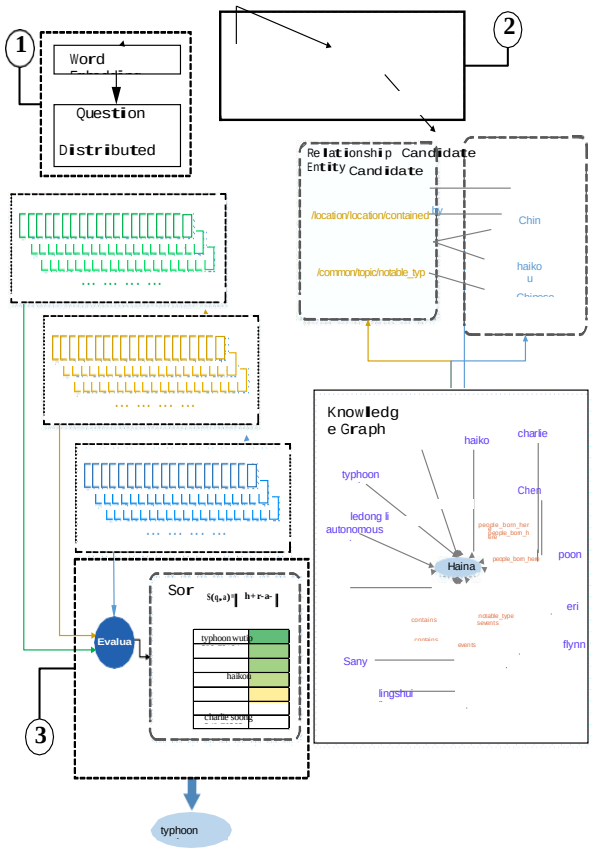

Figure.2. TransE-QA algorithm Sample Overview

6. KNOWLEDGE EXTRACTION AND ATTENTION MECHANISM

The method relating to the knowledge base is introduced, and several works utilize the external knowledge, thereby solving the visual question-answering problem. The focus is on the image domain, and the knowledge video question answering needs to be traversed more. (Jin, Li, Zhao. et al., July 03, 2019) The attempt is to take a step forward in which the method adopts external knowledge extended to the video domain. As depicted in the pictorial representation, Figure. 3. several objects under the video are obtained, serving as the keywords for searching external knowledge. The prevalent object detection task involves detecting a particular class's semantic object instances in videos and images.

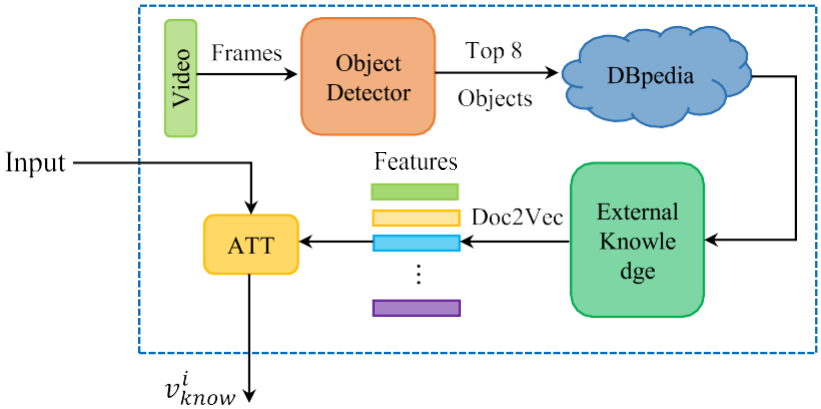

Figure. 3. The Knowledge extraction and attention mechanism architecture

7. CONCLUSION

The grid dataset and NER system produced fills the gap in customer service system technology research in the power grid field. The proposed system has its limitations. The system had to be implemented with high-frequency words, matching the lexicon, and having relatively high data quality requirements in the application field. The BERT model and huge parameters amount influence the system efficiency when deployed on the power grid intelligent customer service system. First, the system efficiency is improved by transferring the learning methodology to reduce the model dependency on the domain data. Secondly, the research proposes simplified BERT models like TinyBERT and MobileBERT. The application of thesimplified BERT models is explored in the system. By considering the intelligent customer service system, both the text information processing and the multimodal information like the sounds and imagesare processed, and future research focuses on the multimodal machine learning technology.

REFERENCES

[1] An-Zi Yen, Hen-Hsen Huang, Hsin-Hsi Chen. CIKM '22: Proceedings of the 31st ACM International Conference on Information & Knowledge Management October 2022 Pages 4645– 4649 https://doi-org.coloradotech.idm.oclc.org/10.1145/3511808.3557717

[2] Shirong Liu, Zixian Guo, Hongzhi Wang. ACM TURC '20: Proceedings of the ACM Turing Celebration Conference - China May 2020 Pages 145–149 https://doi-org.coloradotech.idm.oclc.org/10.1145/3393527.3393552

[3] Yin, Didi; Cheng, Siyuan; Pan, Boxu; Qiao, Yuanyuan; Zhao, Wei; et al. Applied Sciences; Basel Vol. 12, Iss. 11, (2022): 5373. DOI:10.3390/app12115373. Chinese Named Entity Recognition Based on Knowledge-Based Question Answering System. https://www.proquest.com/compscijour/docview/2674337976/793F0680C6E84A42PQ/2?accountid=144789

[4] Yue Wang, Qimai Chen, Chaobo He, Hai Liu, Xiyu Wu. ICIAI 2020: Proceedings of 2020 the 4th International Conference on Innovation in Artificial Intelligence May 2020 Pages 170–179 https://doi-org.coloradotech.idm.oclc.org/10.1145/3390557.3394296

[5] Weike Jin, Zhou Zhao, Yimeng Li, Jie Li, Jun Xiao, Yueting Zhuang. ACM Transactions on Multimedia Computing, Communications, and Applications Volume 15 Issue 2s Article No.: 52. July 03, 2019. pp 1–22 https://doi-org.coloradotech.idm.oclc.org/10.1145/3321505