Knowledge Graph Extraction from the Text

Yuxin He
yhe242@wisc.edu

Yiwu Zhong
yzhong52@wisc.edu

Ching-Wen Wang
cwang553@wisc.edu

1. Introduction

Tremendous information are created and transferred among people everyday. To digest the information from the news, the newspapers and the posts on social media, human naturally grasp the important information, convert them into a few keywords, and finally memorize them. These key information in mind are usually stored in "mind palace". Human can effortlessly query their "mind palace" anytime and find the desired information in seconds. However, for machines, it's not an easy task to process such high volume data in a short time. Knowledge graph, hopefully, an abstract and structured data format, is widely used in industry, such as web search, question answering and data integration. Knowledge graph allows people or machines to quickly digest the key information and to ignore the redundant words in text. If we have a way to create such knowledge graphs from the text we usually meet everyday, it can be beneficial for our work and study efficiency. Now the question is: can we construct a knowledge graph from the text in our daily life?

Knowledge graph is a type of data representation. It consists of a set of nodes that are usually represent the entities (e.g., the nouns) and a set of directed edges among the nodes (e.g., the relationships between nodes). Essentially, knowledge graph can be decomposed into a list of triplets with each as ⟨subject, predicate, object⟩ (e.g., ⟨man, playing, ball⟩). With this structured information, human and people can access the the target node or edge in an efficient way.

The goal of this project is to extract a knowledge graph from the given sentences. Constructing a knowledge graph from given text involves 2 steps: extracting the triplets from each sentence and merging all triplets into the graph format. In this project, we pay more attention to the first step, triplet extraction from text. Since previous models were trained and evaluated in different datasets and settings, it's not sure which model is the suitable one in practice. To this end, we plan to collect our own sentences dataset and label the triplets in sentences. The dataset will be only divided into validation set and test set. After that, we will directly adopt 3 widely-used existing models (Stanford OpenIE [1] and Stanford Scene Graph Parser [5]) or tool (spaCy library[1]) for triplet extraction and evaluate the extracted triplets on the validation set. According to the evaluation results, we select and apply the best method on the test set and obtain a list of triplets. Finally, these extracted triplets form a knowledge graph using lemmatization and our manually-designed rules. The flowchart of our plan is shown in Fig. 1.

For data collection and labeling, we plan to collect the sentences from our daily life, such as newspaper, from image sentence descriptions, or from scientific papers. For each source, we collect a fixed number of sentences (e.g. 100 sentences) and manually label the subject, the predicate and the object in each sentence. These labels are used as ground truth during evaluation. The more the extracted triplets can be matched to these human annotated labels, the better the model is.

The baseline model we plan to exploit is the NLP tool spaCy library, which includes pretrained models and helps users build algorithms that process and understand text. Basically, we first use spaCy library to determine the part-of-speech of each word in the sentences and develop a rule-based strategy to select the subject, the predicate and the object. Since this method relies on the hand-crafted rules, it can be a baseline method compared to other models that were trained by the labeled triplets.

In previous studies, triplet extraction in this project is regarded as information extraction or relation extraction [2, 4, 3, 1, 5]. The existing models we plan to use include Stanford OpenIE [1] and Stanford Scene Graph Parser [5]. Stanford OpenIE [1] first extracts the triplets by recursively traversing the dependency trees that parsed from sentences. And a logistic regression classifier learns to determine which dependencies to be the triplet candidates in a supervised way. Stanford Scene Graph Parser [5] learns to map the dependency syntax representation that parsed from image sentence descriptions to a scene graph, supervised by human annotated scene graphs.

In summary, this project aims at creating a benchmark for the triplet extraction during knowledge graph construction. There are multiple potential ways to evaluate the performance of existing models. In our project, we more care

[1] https://spacy.io/

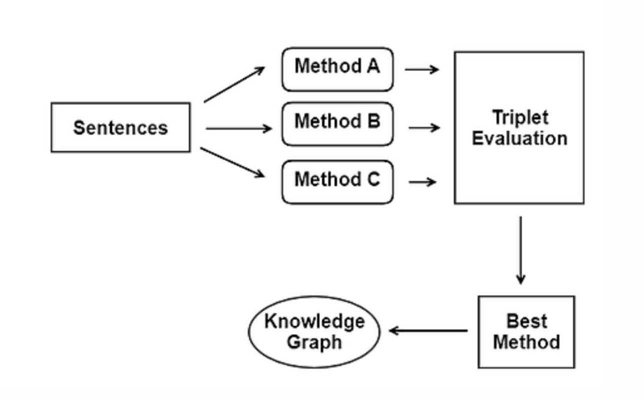

Figure 1. The flowchart of our project pipeline. We first evaluate the quality of triplets that are extracted by different widely-used models. Based on the evaluation results on validation set, we select the best method to apply on the sentences in test set. Finally, a knowledge graph will be constructed by the extracted triplets from the best model.

about which widely used model can be the best practice in real life, such as the collected sentence data that appear in our daily life.

2. Motivation

Humans are capable to interconnect different entities or concepts and understand their relationship automatically. However, distinct from human, computers treat each word independently and interpret them as sequences without any meaning. For example, without the current technology of Natural Language Processing (NLP), a user wants to find information online, but his or her query is structured using imprecise wording, the computer may be unable to deliver relevant information that the user actually needs.

Therefore, if we can utilize technologies that transforms language into real knowledge and apply it to machines, letting computers understand the context of a sentence like human will not merely be a dream anymore. By having a network of knowledge that computers can access to, we can accomplish and improve many inefficiencies in life, such as enhancing user experience in providing better search results or help businesses in creating their own content clusters of clients so that they can optimize any business strategy towards them.

Knowledge graph is one of such technologies. Nowadays, we are seeing the use of knowledge graph in a variety of applications, including web search, answering questions, and data integration. Moreover, knowledge graphs are mentioned in papers regarding Natural Language Processing, computer vision, and other Machine Learning algorithms in a high frequency, where people have started to use knowledge graphs as a way to store extracted information from unstructured text (articles, sentences). Such a process is called information extraction, which processes the extracted structured information for computer understanding.

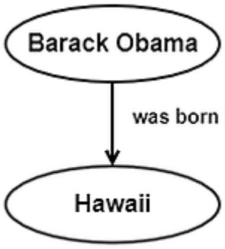

Figure 2. A visualized sample of knowledge graph that extracted from a sentence

In this project, we plan to exploit the models of information extraction to select the key logic and meaning of a set of sentences and generate the corresponding knowledge graph. Most implementations of knowledge graphs use a concept named triplet, that is a tuple of three items (a subject, a predicate, and an object) that we can use to store the key information within a sentence. For example, the raw text is “Barack Obama was born in Hawaii.” After analyzing, the subject is “Barack Obama”, the predicate is “was born in”, and the object is “Hawaii”, as shown in Fig. 2

This project is mainly motivated by the question: can we construct a knowledge graph from the text in our daily life? If we can find a way to create high-quality knowledge graph from the text we usually meet daily, it brings us efficiency for our work and study. The crucial step in knowledge construction is the information extraction. Though major progress has been made in information extraction techniques, in our project, we seek for a best practice among the widely used models so that the best model can be effectively applied on the collected sentence data that appear in our daily life. Hence, we plan to collect our own sentence data from daily news, image sentence descriptions, scientific papers, etc. Then the existing methods are evaluated on the collected data and the best one will be selected to construct the final knowledge graph. The constructed knowledge graph covers the key information of text and thus can be further used for information searching, question query and knowledge archive, which potentially serves as a useful tool in our daily life.

3. Evaluation

A knowledge graph can be regarded as a list of triplets. To evaluate the quality of the output knowledge graph, it’s equivalent to evaluate the quality of extracted triplets.

We plan to use our collected dataset for triplet evaluation. The dataset consists of a set of sentences collected by our own. For example, we can collect the sentences from daily newspaper, from image sentence descriptions or from scientific papers. A fixed number of sentences (e.g. 100 sentences) are collected from each source. Then these sentences are manually labeled with the subject, the predicate and the object.

The metrics we will use include: precision, recall and F1 score. Specially, these metrics can be evaluated on different aspects of triplets. We can only evaluate the extracted subjects/objects compared to the labels. Besides, the extracted triplets are treated as correct cases if they are successfully matched to the labeled triplets as a whole.

The pipeline in this project involves the essential components in machine learning, including: data collection and labeling, model evaluation and selection, and multiple fundamental areas in NLP (part-of-speech tagging, sentence dependency tree parsing, lemmatization and stemming, and information extraction / relation extraction). If we successfully construct the whole pipeline, evaluate the extracted triplets in a reasonable way and the final output knowledge graph is able to represent the important concepts from the input sentences, our project can be treated as ”successful”.

4. Resources

The resources we are going to used are sumarized as follows:

- spaCy library: a widely used tool in NLP
- Stanford OpenIE: information extraction tool [1]
- Stanford Scene Graph Parser: parses a sentence into triplets [5]

5. Contributions

This project can be roughly divided into 3 parts: data preparation, model evaluation and result analysis. We plan to distribute the work in each apart evenly to each team member. For example, each person can collect sentences from a different data source, run and evaluate one of the models we plan to use. Finally, we will analyze the results and write the report together.

References

[1] G. Angeli, M. J. J. Premkumar, and C. D. Manning. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, 2015.

[2] B. Distiawan, G. Weikum, J. Qi, and R. Zhang. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 229–240, 2019.

[3] J. Kuang, Y. Cao, J. Zheng, X. He, M. Gao, and A. Zhou. Improving neural relation extraction with implicit mutual relations. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1021–1032. IEEE, 2020.

[4] J. L. Martinez-Rodriguez, I. Lopez-Arevalo, and A. B. Rios-Alvarado. Openie-based approach for knowledge graph construction from text. *Expert Systems with Applications*, 113:339–355, 2018.

[5] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015.