Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 7, No. 2, 2024

Information Extraction and Knowledge Map Construction based on Natural Language Processing

Zehan Wang

University of Maryland, College Park, Maryland, USA

Abstract: As a key branch of artificial intelligence, Natural Language Processing (NLP) focuses on making machines understand and generate human language. This paper introduces the basic tasks of NLP, such as lexical analysis, syntactic analysis and semantic understanding, and discusses the cutting-edge technologies such as word embedding. In the aspect of information extraction, this paper deeply discusses the methods of named entity recognition, relationship extraction and event extraction, and points out the challenges in dealing with complex texts. Finally, the paper focuses on the construction of knowledge map, expounds the complete process from data collection to entity identification, relationship extraction, graph construction and query, and emphasizes the core position of graph query in the application of knowledge map. On the whole, this paper provides a comprehensive perspective for understanding NLP, information extraction and knowledge map construction, and points out the importance and future development direction of these technologies in intelligent systems.

Keywords: Natural Language Processing; Information Extraction; Knowledge Map.

1. Introduction

As information technology rapidly advances, the volume of text data on the Internet explodes, encompassing vast amounts of knowledge. The challenge of efficiently extracting valuable insights from this data and constructing a structured knowledge system has emerged as a significant focus of contemporary research [1]. The technology of information extraction and knowledge map construction based on NLP came into being, which provided new ideas and methods for solving this problem. NLP is an important branch in the field of artificial intelligence, which aims to enable computers to understand and process human language. As one of the core tasks of NLP, information extraction aims to extract structured information from unstructured text data, such as entities, relationships, events and so on [2]. Through information extraction technology, the knowledge points scattered in the text can be extracted, which provides a data basis for the subsequent knowledge map construction.

Knowledge map is a technology to represent knowledge by graph structure, which represents and stores entities, concepts and their relationships in the real world in a structured form [3]. Through the knowledge map, we can understand and describe various concepts and relationships in the real world more clearly and realize the sharing and reuse of knowledge [4]. Knowledge map has a wide application prospect in intelligent question answering, semantic search, intelligent recommendation and other fields.

This paper aims to study the technology of information extraction and knowledge map construction based on NLP. Firstly, the basic theory and key technologies of NLP are introduced to provide theoretical support for subsequent information extraction and knowledge map construction. Then, the main methods and implementation technologies of information extraction are discussed, including named entity recognition, relationship extraction and event extraction. On this basis, we will further study the key technologies such as the construction process of knowledge map, knowledge representation and modeling methods, entity linking and disambiguation.

The research of this paper has important theoretical significance and practical value. In theory, through the in-depth study of NLP and information extraction technology, we can further improve and develop the relevant theoretical system and provide new ideas and methods for the follow-up research. In practice, by constructing high-quality knowledge map, we can provide more abundant and accurate knowledge support for intelligent question answering, semantic search and other application fields, and promote the further development of artificial intelligence technology.

2. NLP and Information Extraction

NLP is an interdisciplinary field of artificial intelligence (AI) and linguistics, aiming at making machines understand and generate human language. With the rise of deep learning, NLP has made remarkable progress, especially in text classification, sentiment analysis, machine translation and so on. As an important branch of NLP, information extraction focuses on extracting structured information from unstructured texts, providing data support for building knowledge maps and intelligent question answering systems [5].

(1)NLP foundation

NLP involves many tasks, including lexical analysis, syntactic analysis and semantic understanding. Lexical analysis mainly deals with lexical problems, such as word segmentation and part-of-speech tagging [6]. Syntactic analysis focuses on the structural relationship between words in a sentence, such as dependency syntax analysis and phrase structure analysis. Semantic understanding aims to capture the deep meaning of the text, including named entity recognition, emotion analysis, semantic role labeling and so on.

In NLP, text is usually represented as a vector or matrix so that the machine learning model can handle it. Bag model and TF-IDF (TERM frequency-inverse document frequency) are two common text representation methods. The bag-of-words model regards text as a collection of words, ignoring the order and grammatical structure between words. TF-IDF measures the importance of vocabulary by considering the frequency of

vocabulary in the document and the frequency of reverse document.

(2) Information extraction technology and method

The main task of information extraction is to extract structured information units from unstructured texts, such as entities, relationships, events, etc. [7]. Named Entity Recognition (NER) is a key step in information extraction, which aims to identify entities with specific meanings in texts, such as names of people, places and organizations. NER is usually implemented by rule-based, dictionary-based or machine learning-based methods. With the development of deep learning, neural network-based methods, such as BiLSTM-CRF model, have achieved remarkable performance improvement on NER tasks.

Relationship extraction is another important task of information extraction, which aims to identify the relationship between entities in the text. Relationship extraction can be realized by rule-based, template-based or machine learning. In recent years, relationship extraction methods based on deep learning, such as PCNN (Piecewise Convective Neural Networks) and Attention mechanism, have achieved excellent performance in the task of relationship extraction. In addition, joint extraction models such as CASREL (Cascade Relation Extraction) can identify entities and relationships at the same time, which improves the accuracy and efficiency of extraction.

(3) The practical application and challenge of information extraction.

Information extraction technology is widely used in many fields, such as intelligent question answering, semantic search, bioinformatics, enterprise knowledge management, etc. In intelligent question answering system, information extraction technology can extract answers or evidence related to questions from a large number of documents; In semantic search, information extraction technology can help users find the information they need more accurately; In bioinformatics, information extraction technology can extract knowledge such as protein relation and gene function from biomedical literature. In enterprise knowledge management, information extraction technology can help enterprises build internal knowledge system and improve the work efficiency of employees.

However, information extraction technology also faces some challenges and problems. First of all, how to improve the accuracy and efficiency of extraction when dealing with complex texts is a key issue. To solve this problem, researchers put forward various optimization methods and techniques, such as using deep learning model to capture the deep features in the text and using attention mechanism to pay attention to important information. Secondly, how to effectively use external knowledge and prior information to improve the extraction performance is also a problem worth studying. In order to solve this problem, researchers try to introduce external knowledge base or pre-trained language model into the information extraction process to provide richer semantic information. Finally, how to build a large-scale and high-quality training data set is also a challenging problem.

3. Knowledge Map Construction

Knowledge map is a semantic network that graphically represents entities and their relationships. By constructing knowledge map, we can systematically organize, understand and share domain knowledge, and then support advanced applications such as intelligent question answering, recommendation system and semantic search. This section will introduce the construction process of knowledge map in detail, and interspersed with relevant formulas to illustrate the key technologies.

(1) Construction process of knowledge map

The construction of knowledge map usually includes the following steps: data collection and preprocessing, entity identification and linking, relationship extraction, map construction and query [9]. In the data collection stage, it is necessary to obtain the original data from various sources (such as text, database, images, etc.) and carry out preprocessing operations such as cleaning, deduplication and formatting. Entity identification and linking is one of the key steps in the construction of knowledge map, aiming at identifying entities with specific meanings from texts and linking them to corresponding knowledge bases or ontologies. Relationship extraction focuses on extracting the relationship between entities from the text to form a structured knowledge representation. Finally, in the graph construction and query stage, the extracted entities and relationships are integrated into a graphical data structure, and an efficient query and reasoning mechanism is provided.

(2) Entity identification and linking

Entity recognition is one of the basic tasks of knowledge map construction, and its goal is to identify named entities (such as names of people, places, organizations, etc.) in texts, and to classify and standardize them. Common entity recognition methods include rule-based method, statistical method and deep learning method [10]. Rule-based methods usually rely on hand-written rules and dictionaries, which are suitable for entity recognition tasks in specific fields or scenarios. The method based on statistics uses machine learning algorithm to learn the model of entity recognition from labeled data, which has good universality and expansibility. Deep learning method automatically extracts text features and classifies them through neural network model, and has made remarkable progress in entity recognition tasks in recent years.

Entity linking is the process of linking the identified entity to the corresponding concept or entity in the knowledge base or ontology. The key of entity linking is to calculate the similarity or matching degree between entity reference items and candidate entities in knowledge base. Commonly used similarity calculation methods include string-based method, semantic-based method and graph-based method. The string-based method mainly uses the string matching algorithm to calculate the similarity between the reference item and the candidate entity. Semantic-based methods use techniques such as word embedding and semantic role labeling to capture the semantic relationship between referents and candidate entities. The graph-based method uses the graph structure information in the knowledge map to calculate the correlation between the reference item and the candidate entity.

Entity link can associate entities in the text with those in the knowledge map. This paper proposes to use a simple neural network model to predict whether two entities belong to the same entity. The model can be expressed as:

$$P(e_1 = e_2) = \sigma(W_1 \cdot (e_1 + e_2) + b_1) \quad (1)$$

Where e_1 and e_2 represent two entity vectors respectively, and W_1 and b_1 are model parameters. Function σ is an activation function, such as ReLU or

sigmoid function.

In the construction of knowledge map, entity disambiguation is an important problem, which aims to identify the entities in the text uniquely. Similarity-based methods can be used to measure the similarity between entities and determine whether they are the same entity. Commonly used similarity indicators include cosine similarity and Euclidean distance. For example, cosine similarity can be used to calculate the similarity between entities:

$$\text{CosineSimilarity}(e_1, e_2) = \frac{e_1 \cdot e_2}{\|e_1\| \|e_2\|} \quad (2)$$

Where e_1 and e_2 respectively represent two entity vectors, $e_1 \cdot e_2$ represents the dot product of the two vectors, and $\|e_1\|$ and $\|e_2\|$ represent the norms of the two vectors.

(3) Relationship extraction

Relationship extraction is another key task in the construction of knowledge map, which aims to extract the relationships between entities from the text and form a structured knowledge representation. Relationship extraction can adopt rule-based method, template-based method or deep learning-based method. Rule-based methods usually rely on relationship extraction rules defined by domain experts, and are suitable for relationship extraction tasks in specific fields or scenarios. The template-based method uses predefined templates to match the relational patterns in the text, which has good universality and expansibility. Deep learning method automatically extracts text features and classifies relationships through neural network model, and has made remarkable progress in the task of relationship extraction in recent years.

For each pair of possible entities, a classifier is used to judge whether there is a certain relationship between them:

$$SVM(f(x), y) = \sum_i \alpha_i y_i (f(x) - y_i) + b \quad (3)$$

Where $f(x)$ is the input feature vector, y is the label (positive or negative), α_i is the Lagrangian multiplier, and b is the offset term.

For entities that do not appear in the knowledge map, they need to be associated with the entities in the knowledge map through entity linking technology:

$$sim(e_{text}, e_{kg}) = \frac{2}{1/d(e_{text}) + 1/d(e_{kg})} \quad (4)$$

Among them, e_{text} is the entity in the text, e_{kg} is the entity in the knowledge map, and d is the similarity function of the entity description.

4. Conclusion

With the continuous progress of artificial intelligence technology, NLP and information extraction, knowledge map construction and other fields are increasingly in-depth. This paper discusses the basic technology of NLP, the key method of information extraction, and the process and core technology of knowledge map construction. In NLP, the development of word embedding technology provides a new idea for the semantic representation of words, which enables machines to better understand and process human language. As an important branch of NLP, information extraction, with its core technologies such as named entity recognition, relationship extraction and event extraction, can effectively extract structured information from unstructured texts, which provides strong support for building large-scale knowledge maps.

The construction of knowledge map is a comprehensive process, involving data collection, entity identification and linking, relationship extraction and map construction. This paper introduces the key technologies and methods of these links, and emphasizes the importance of entity identification and linking and relationship extraction in the construction of knowledge map. The continuous development of graph construction and query technology makes the application of knowledge map more extensive and efficient.

The research in NLP, information extraction and knowledge map construction not only promotes the progress of artificial intelligence technology, but also provides strong support for many application scenarios such as intelligent question answering, semantic search and enterprise knowledge management. In the future, with the continuous innovation and development of technology, the research in these fields will continue to deepen, bringing more convenience and possibility to human life and work.

References

[1] Rhinehart R R. An introduction to natural language processing PART 2[J]. Control, 2021, 2021(4):34.

[2] Jeon J H, Xu X, Zhang Y, et al. Extraction of Construction Quality Requirements from Textual Specifications via Natural Language Processing:[J]. Transportation Research Record, 2021, 2675(9):222-237.

[3] Han X, Zhang Z, Liu Z. Knowledgeable Machine Learning for Natural Language Processing[J]. Communications of the ACM, 2021, 2021(11):64.

[4] Aldabbas H, Bajahzar A, Alruily M, et al. Google Play Content Scraping and Knowledge Engineering using Natural Language Processing Techniques with the Analysis of User Reviews[J]. Journal of Intelligent Systems, 2020, 30(1):192-208.

[5] Pham H T T L, Han S U. Natural Language Processing with Multitask Classification for Semantic Prediction of Risk-Handling Actions in Construction Contracts[J]. Journal of computing in civil engineering, 2023, 2023(6):37.

[6] Tsai M F, Tseng H J. Enhancing the identification accuracy of deep learning object detection using natural language processing[J]. The Journal of Supercomputing, 2021, 77(7):1-16.

[7] Liu L, Yu Q. Research on classification method of answering questions in network classroom based on natural language processing technology[J]. International journal of continuing engineering education and life-long learning, 2021, 2021(2):31.

[8] Zhang C, Mayr P, Lu W, et al. Guest editorial: Extraction and evaluation of knowledge entities in the age of artificial intelligence[J]. Aslib Journal of Information Management, 2023, 75(3):433-437.

[9] Sawant A A, Devanbu P. Naturally!: How Breakthroughs in Natural Language Processing Can Dramatically Help Developers[J]. IEEE Software, 2021, 38(5):118-123.

[10] Mustafa H A, Al-Wesabi F N, Abdelzahir A, et al. A Hybrid Intelligent Text Watermarking and Natural Language Processing Approach for Transferring and Receiving an Authentic English Text Via Internet[J]. The Computer Journal, 2021, 2021(2):2.