Architecture of Knowledge Extraction System based on NLP

Wei Zhuang*
Pactera EDGE, Redmond, USA, e-mail:
Zhuangwei2021@126.com

ABSTRACT

Knowledge extraction is to extract useful structured text information from messy free text. Under the current massive information background, it has attracted extensive attention. This paper analyzes the concept of NLP and the application process of NLP algorithm, discusses web information retrieval system, information extraction based on natural language processing and text relationship extraction, and tests the pipeline performance. The results show that the pipeline time is not linearly correlated with the size of the novel, but positively correlated.

CCS CONCEPTS

• **Applied computing** → Physical sciences and engineering; Engineering.

KEYWORDS

NLP, Knowledge Extraction System, NLP Algorithm, Text Relation Extraction

ACM Reference Format:
Wei Zhuang. 2021. Architecture of Knowledge Extraction System based on NLP. In *2021 International Conference on Aviation Safety and Information Technology (ICASIT 2021), December 18–20, 2021, Changsha, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3510858.3510950

1 INTRODUCTION

With the rapid development of Internet technology, more and more web pages are published, and a large amount of information appears in front of us in the form of electronic documents. People urgently need to use some automatic tools to eliminate the thickness, false, retain the truth, and quickly find the valuable information they need.

With the continuous development of science and technology, many experts have studied NLP knowledge extraction system. For example, noura m, gyrard a, Heil s designed knowledge extraction for wot (ke4wot) using domain specific knowledge encoded in Internet of things publications. The descriptions of these topics of the ten most popular ontologies in the three fields are compared with the experience evaluation of 23 domain experts. The results show that the topic of Internet of things ontology can be used as a keyword to fully describe the existing ontology [1]. Hou J, Li x, Yao h proposed a transformer bidirectional coded representation (BERT) based on public security Chinese relation extraction algorithm, which can effectively mine security information. The main model structure is laminated transformer [2]. Majid a, Andrew P, niranjan k the purpose of the study was to train and validate the NLP classifier for identifying patients with alcohol abuse. External validation is required before it is applied to enhanced screening [3]. Although the research results of NLP knowledge extraction system are quite rich, the research on the architecture of knowledge extraction system based on NLP is still insufficient.

In order to study the architecture of natural language processing knowledge extraction system, this paper studies natural language processing and knowledge extraction system, and finds probabilistic soft logic. The results show that NLP algorithm is helpful to build the architecture of knowledge extraction system.

2 METHOD

2.1 NLP

2.1.1 Concepts of natural language processing. Natural language processing is a research direction in the field of computer science and engineering. The application fields of natural language processing mainly include speech recognition, document summarization, document classification, etc. [4]. However, it is not feasible to just input a large amount of data into the computer in the hope that it can learn to speak. We should first prepare data that is convenient for computers to discover patterns and reasoning, and then achieve this goal by adding relevant metadata to the data set. It can be seen that the research of corpus is the key link in the development of intelligent human solving technology [5]. Syntactic analysis is one of the key technologies of natural language processing (NLP). Its purpose is to determine the syntactic structure of sentences or the dependence between sentence components. Dependency analysis usually uses dependency tree to graphically display the syntactic structure of sentences, that is, the linear relationship between word pairs is hierarchized into tree structure [6].

2.1.2 NLP algorithm application process. The whole process of NLP algorithm consists of multiple interrelated stages [7]. The general process is lexical analysis → syntactic analysis → semantic analysis → pragmatic analysis → text analysis. However, for short texts, that is, texts without paragraphs or whole articles, pragmatic analysis and text analysis are meaningless. It mainly analyzes the correlation between paragraphs and context in long texts. The main purpose of lexical analysis is to segment the original input sentence into strings through word segmentation algorithm, and label the part of speech of the corresponding strings to form a group of labeled strings; The main purpose of syntactic analysis is to master the part of speech correspondence between word strings according to the part of speech set marked, and finally form the corresponding semantic

[*] Corresponding author

ICASIT 2021, December 18–20, 2021, Changsha, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-9042-2/21/12…$15.00
https://doi.org/10.1145/3510858.3510950

set; The main purpose of semantic analysis is to understand the overall meaning of the sentence after obtaining the above analysis results [8]. However, from the perspective of core algorithms, the main difference is that different languages use NLP technology toolkit to construct recognition rules or patterns according to the grammatical characteristics of corresponding languages of different languages. Generally speaking, most automatic generation tools of UML models are realized through NLP technology combined with rule matching or pattern recognition [9].

2.2 Knowledge Extraction System

2.2.1 Network information retrieval system. Search engine is one of the most common web information retrieval systems, such as Google and Alta vista. The flexibility of HTML makes the structure of web pages not benign. For example, some statements have only a start tag but no end tag. When we extract information according to the inherent semi-structured structure of web pages, this non benign structure will bring us a lot of trouble. Therefore, considering that there are still a large number of HTML pages and web applications on the Internet, we first need to convert non benign web pages into benign web pages. Because the document generated using XML has a good structure, and XHTML is an extension of HTML and is based on XML, we can convert the original HTML document into XHTML document. Here, we use tidy to "fix" non benign tags in HTML documents [10]. Compare with the web page function or link information in the knowledge base, apply the closest knowledge to the web page, automatically extract the information in the web page and feed it back to the user, or display the information of interest to the user in a striking way. In the future interaction with users, intelligence can modify existing knowledge or add new knowledge by recording users' feedback, so as to continuously improve the accuracy of information query.

2.2.2 Information extraction based on natural language processing. Information extraction technology based on natural language processing is suitable for documents containing a large number of text. It is mainly used to extract data from free text. The input is a document and a complete template indicating the data to be extracted. The first simulation test template is used to extract data extraction patterns and extract data by instantiation [11]. The learning algorithm combines the technology involved in some logic programming tools and unbounded pattern learning technology. It learns and extracts rules from text documents according to a given training sample set. It relies on a token oriented feature set. These functions can be simple or related. In order to better analyze the data, you need to extract the text data in the XML file and save it as a TXT file. It is the preliminary preparation for other natural language processing tasks, such as information extraction, knowledge map construction and so on. It usually refers to identifying named entities such as person name, name, proper noun and organization name, but it will also identify different types of entities in different fields according to the characteristics of different fields.

2.2.3 Text relation extraction. Text relation extraction is an important part of information extraction. Its main task is to give the relationship to be extracted and identify the sentences describing the relationship from the text. Similar to other classification problems, text-based relationship extraction can use knowledge mapping to train the model. However, this idea requires manual construction of large-scale knowledge map, which requires not only experts with professional skills, but also a large number of labor force. Machine learning method can overcome the shortcomings of knowledge mapping method. This method does not need to be built manually by experts with professional skills. It only needs people with certain professional knowledge to judge whether the relationship between any two entities is the relationship we need [12].

2.3 Probabilistic Soft Logic

Probabilistic soft logic has been widely used in the fields of collective classification, social trust analysis, personalized recommendation and so on. Probabilistic soft logic inference rules are composed of weighted first-order logic rules, as shown in formula (1):

$$friend\,(x, y) \wedge voteFor\,(y, z) \Rightarrow voteFor\,(x, z) \tag{1}$$

PSL rules show that if x and y are friends and Y votes for Z, then x also has a certain probability to vote for Z, and the voting probability is represented by the weight W.

Logical predicates have their own interpretation probability, indicating the possibility of something happening. The probability of satisfying rule R is $\varphi(R)$, calculated according to formula (2):

$$\varphi\,(r) = \max\left\{0, \left(I_{body} - I_{head}\right)\right\} \tag{2}$$

In equation 2),I_{body} and I_{head} represent the probability values of rule body and rule header respectively. The digital operation method of logical symbols is shown in equation 3):

$$I\,(p \wedge q) = \max\,(0, p + q - 1) \tag{3}$$

P and Q represent logical predicates.

3 EXPERIENCE

3.1 Object Extraction

In the e-commerce information extraction system based on knowledge map, the whole system is divided into three layers: interaction layer, business logic layer and storage layer. The interaction layer of the system is responsible for the input and output of the whole system, and transmits the system information to users through web pages. The interaction layer consists of two page views: task view and extraction result view. In the task view, you can generate new tasks and view the information of all tasks, while in the extraction results view, you can view the extraction results of tasks. The knowledge map service module is responsible for all operations of the knowledge map in the system. The module is composed of knowledge map basic service module and knowledge map attribute domain value service module. The basic service module of knowledge map takes the knowledge map of Chinese encyclopedia as the data source, reconstructs the storage form according to the system requirements, constructs the index for the map content, and provides external storage and query services for the knowledge map.

Table 1: Test comparison of optimization algorithms

Optimization algorithm	Accuracy	F value
Random gradient descent algorithm	26%	12%
Adaptive gradient algorithm	34%	34%
Root mean square algorithm	53%	54%
Adaptive estimation algorithm	56%	53%

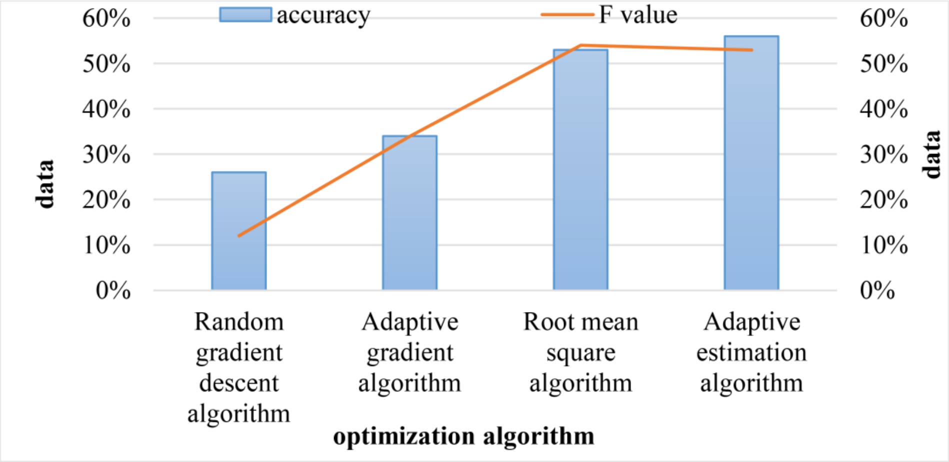

Figure 1: Test comparison of optimization algorithms

3.2 Experimental Analysis

Step 1: table identification. Table recognition refers to finding the table area of the domain data to be extracted from the web page and removing noise such as "false tables" (such as advertisements and navigation bars). The second step is to standardize web tables. The cells in the web page table occupy multiple rows (columns), so you need to align each row (column) cell in the table with the same number. Step 3: table structure identification. According to the structure type of the table and the expansion method of attribute value pairs, it can be divided into horizontal and vertical types, also known as rows and columns. In this module, the expansion mode of the table and the location of table attribute rows (columns) and data cells are determined. Step 4: table content extraction. The early developed steel material semantic model STSM obtains the attributes in the ontology (object attributes and data attributes) and the attributes in the table for string matching. If the matching result is greater than a threshold, the attribute value is extracted. Fifth, generate RDF data. The RDF chain dataset is generated using the data extracted from the table and the matching attributes in the STSM ontology.

4 DISCUSSION

4.1 Influence of Optimization Algorithm on Model

In the research of neural network word segmentation algorithm, selecting the appropriate optimization algorithm is of great significance to the training model and model optimization. In essence, the optimization algorithm adjusts the loss function by adjusting the training mode to make the training result of the model close to the ideal state. The most intuitive negative impact on the selection of optimization algorithm is that the training results of the model will reach the local optimal state rather than the global optimal state (see table 1).

It can be seen from the above that the accuracy of the random gradient descent algorithm is 26% and the F value is 12%; The accuracy of adaptive gradient algorithm is 34%, and the F value is 34%; The accuracy of root mean square algorithm is 53%, and the F value is 54%; The accuracy of the adaptive estimation algorithm is 56%, and the F value is 53%; The specific presentation results are shown in Figure 1

Stochastic gradient descent algorithm and adaptive gradient algorithm do not converge under the same number of iterations because they are not optimized in the model. Adaptive estimation algorithm is the optimization algorithm of root mean square algorithm, so its actual performance is slightly better than root mean square algorithm. Therefore, in the experiment, the adaptive estimation algorithm is selected as the optimization algorithm of the model.

4.2 Performance Test

As an offline task, the system designed in this paper does not require high timeliness and resource consumption, but this section will still evaluate the time-consuming of the system. The system selects three different types of novels: martial arts novel Tianlong Babu, white horse roaring west wind, romantic novel Hello, Mr. Jin and online game novel full-time Master. In addition to different types

Table 2: System performance test

Novel	Size(KB)	Time consuming(s)
Tianlong Babu	4256	122
White horse whistling west wind	312	42
Full time expert	532	35
Hello, Mr King	257	27

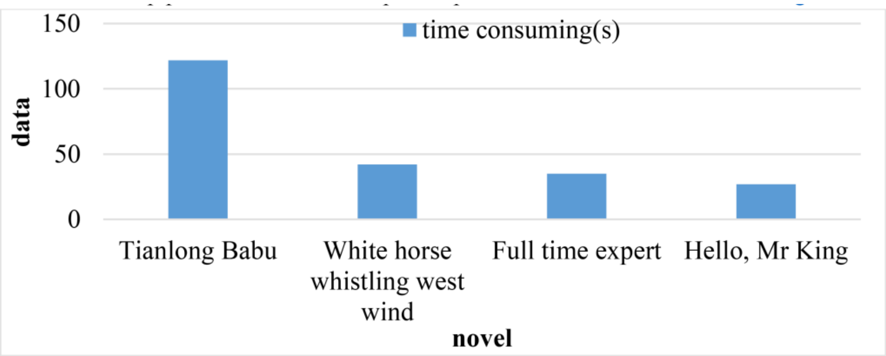

Figure 2: System performance test

and styles, the number and scale of words in the three novels are also different. In this section, the overall pipeline test of three novels is carried out on K20 GPU under linux environment. The test results are shown in Table 2

It can be seen from the above that the file size of Tianlong Babu is 4256kb, the pipeline takes 122s, the file size of white horse roaring west wind is 312kb, the pipeline takes 42s, the file size of full-time Master is 532kb, the pipeline takes 35S, the file size of Hello, Mr. king is 257kb, and the pipeline takes 27s. The specific presentation results are shown in Figure 2

It can be seen from the above that the pipeline time consumption is not linearly correlated with the new size, but positively correlated. And the time-consuming of the system is acceptable. Generally speaking, a novel can be predicted in a few minutes. For the longer novel Tianlong Babu, it can be predicted within 5 minutes in the case of single machine and single card. Compared with manual reading, the efficiency is greatly improved, and the whole system is stable and reliable.

5 CONCLUSION

With the rapid development of the Internet, the information on the Internet is growing explosively. The emergence and popularity of the world wide web and web browsers make it easier and more intuitive for users to obtain information, and also makes it possible for search engines such as Google, which further reduces the difficulty for users to obtain information. This paper studies the influence of optimization algorithm on the model. The results show that the stochastic gradient descent algorithm and the adaptive gradient algorithm do not converge under the same number of iterations because they are not optimized in the model.

REFERENCES

[1] Noura M., Gyrard A., Heil S., *et al.* Automatic Knowledge Extraction to build Semantic Web of Things Applications. IEEE Internet of Things Journal, 2019, 6(5):8447-8454.

[2] Hou J., Li X., Yao H., *et al.* BERT Based Chinese Relation Extraction for Public Security. IEEE Access, 2020, PP(99):1-1.

[3] Majid A., Andrew P., Niranjan K., *et al.* Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. Journal of the American Medical Informatics Association, 2019(3):254-261.

[4] Chen R., Ho J C., Lin J. Extracting medication information from unstructured public health data: a demonstration on data from population-based and tertiary-based samples. BMC Medical Research Methodology, 2020, 20(1):1-11.

[5] M Müller, Alexandi E., Metternich J. Digital shop floor management enhanced by natural language processing. Procedia CIRP, 2021, 96(7-8):21-26.

[6] Zhu H., He C., Fang Y., *et al.* Fine Grained Named Entity Recognition via Seq2seq Framework. IEEE Access, 2020, PP(99):1-1.

[7] Ataeva O M., Serebryakov V A., Tuchkova N P. Ontological Approach: Knowledge Representation and Knowledge Extraction. Lobachevskii Journal of Mathematics, 2020, 41(10):1938-1948.

[8] Taskin Z., Al U. Natural Language Processing Applications in Library and Information Science. Online Information Review, 2019, ahead-of-print(4):676-690.

[9] Guo Q., Qiu X., Xue X., *et al.* Low-Rank and Locality Constrained Self-Attention for Sequence Modeling. IEEE / ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(12):2213-2222.

[10] Tebaldi M., Calaresu M., Purpura A. The power of the President: a quantitative narrative analysis of the Diary of an Italian head of state (2006–2013) . Quality & Quantity, 2019, 53(6):3063-3095.

[11] Myagmar B., Li J., Kimura S. Cross-Domain Sentiment Classification With Bidirectional Contextualized Transformer Language Models. IEEE Access, 2019, PP(99):1-1.

[12] Liu G., Liu D. Treatment of efficiency for temperature and concentration profiles reconstruction of soot and metal-oxide nanoparticles in nanofluid fuel flames. International Journal of Heat and Mass Transfer, 2019, 133(APR.):494-499.