Research on the Application of NLP Artificial Intelligence Tools in University Natural Language Processing

Aihong Yuan1,*, li Gao2

1School of Foreign Languages Xi’an Shiyou University Xi’an, P. R. China
2School of Electronic Engineering Xi’an Shiyou University Xi’an, P. R. China

*Corresponding author: hyuan@xsyu.edu.cn

Abstract. Natural language formal analysis theory has created brilliant achievements on the basis of previous studies. However, with the development of computing power and the advent of the deep learning boom, some people believe that the rule-based rationalist method is outdated, and deep learning that relies on massive data can truly realize artificial intelligence. When traditional natural language is directly transplanted to text language, the short content of natural language will cause data sparseness and result in deviation of calculation results. This paper proposes a new natural language similarity measurement method by using NLP artificial intelligence tools. This method first preprocesses short texts, then builds a complex network model for natural language, calculates the complex network feature values of natural language words, and then uses NLP artificial intelligence tools to calculate the semantic similarity between natural language words, and then combines natural language Semantic similarity is defined to calculate the similarity between natural languages.

Keywords: NLP artificial intelligence tools; university natural language; short text; text similarity rate.

1. Introduction
With the rapid development of the times, intelligent technology continues to advance, intelligent products emerge in endlessly, and people's lives are becoming more and more convenient. The development of artificial intelligence (AI) technology is inseparable from knowledge. Knowledge is the correct conclusion obtained by the information receiver through the refining and reasoning of the information. It is the knowledge and mastery of the natural world, human society, and the way of thinking and movement rules. It is the information that the human brain reassembles and systemizes through thinking, set. The use of knowledge by artificial intelligence technology mainly relies on the knowledge base. The knowledge base (KB) aims to organize human knowledge in a structured form and is playing an increasingly important role in the field of big data and artificial intelligence.

Natural language processing (NLP) is known as "the jewel in the crown of artificial intelligence" and has now become one of the core areas of artificial intelligence. The construction of language knowledge base is a very important applied basic research in the field of natural language processing. Countries all over the world are committed to the construction of large-scale semantic dictionaries and semantic knowledge bases. For example, the English Word Network of Princeton University is a
large-scale machine-readable dictionary incorporating knowledge. The most widely used knowledge base in the field of NLP today is Word Net [1]. Since the release of Word Net version 1.0 in July 1991, the academic community has not interrupted its research. Subsequently, Word Net versions have been continuously updated, and various applications have emerged in an endless stream. It has shown strong vitality in the fields of AI and NLP. In recent years, countless people have devoted themselves to the research and use of Word Net, and they also have a large amount of financial support. Its popularity is getting higher and higher. In today’s various international academic conferences in the field of AI and NLP, the development and use of Word Net and other knowledge bases have attracted much attention in academia and industry. Compared with the study of English, the study of Chinese in the field of NLP has not received enough attention, and the development of the Chinese knowledge base is also affected. In this context, in order for AI and NLP technology to achieve rapid development in the Chinese language field, the Chinese knowledge base needs the attention, support, and research application of academic and industrial circles.

2. Theoretical knowledge

2.1. Related research on Yiyuan knowledge in HowNet

Specifically, a word in HowNet may have multiple meanings, representing the semantic meaning of the word in reality. Each meaning is defined as a series of hierarchical structure of meanings, the structure is shown in the dashed box on the left side of Figure 1. HowNet has unparalleled advantages in integrating deep learning. Different from the organizational model of knowledge bases such as Word Net and Synonyms Calin, HowNet directly and accurately portrays semantic information through a unified meaning original labeling system. The meaning of each meaning is clear and fixed, and can be used directly as a semantic label into a machine learning model. Therefore, the knowledge base can be easily applied to downstream NLP tasks.

![Figure 1. The tree structure of Yoshihara](image)

2.2. Research on the feature extraction method based on the definition information of deep learning

To use the definition or description information of a word to predict the meaning of the target word, the first thing is to study how to extract the characteristics of the definition or description information. Using deep learning methods can avoid complex data processing and feature engineering. This topic uses representation learning technology to first obtain the representation of the word, and then obtain the representation of the sentence through the encoder, which is the feature of the definition information. Specifically, this topic is based on the principle of language model, using convolutional neural network CNN and recurrent neural network such as LSTM and its variants to design the
encoder, compare the advantages and disadvantages of each model through a simple classification task, and select the best performance. The neural network model, as the feature extraction module of Liyuan prediction, solves the problem of the adequacy of feature extraction [2].

3. Yoshihara prediction algorithm

In order to unify the symbolic expression of the formula, first introduce related terms and their symbolic expressions. This article refers to the defined word in the dictionary as the target word, which is different from the word in the definition. The set of words is represented by \( w \). The meaning of the target word is the object of prediction, and \( S \) is used to represent the set of meanings. Given a word \( \omega \in W \), its definition is expressed as \( D_\omega = (d_1, d_2, ..., d_{|D_\omega|}) \), where \( d_i \) is the \( i \)-th word in the definition, and \( |D_\omega| \) is the number of elements in the set. The original set \( S_\omega = (s_{1}, s_{2}, ..., s_{k}) \subseteq S \) of the word \( \omega \). The bold type corresponding to each symbol indicates the corresponding vector.

The collaborative filtering method was originally used in the user recommendation system. It is a method of recommending new targets for new users based on the preferences of similar users based on the principle of similar users. In the semantics prediction task, based on the same principle, similar words have similar semantics, and unlike in the user recommendation task, the subject is human and the uncertainty is large. The meaning of the word is based on linguistics. The principle is marked and conforms to certain specifications, so the noise contained is relatively small, and the collaborative filtering method is more suitable. In SPWE, the synonymous prediction method based on collaborative filtering, the similarity between words uses the cosine distance between word vectors. Formula (1) defines the formula for Yoshihara’s prediction score:

\[
P(s, w) = \sum_{w' \in w} \cos(w, w') M_{y}c^{r_i}
\]  

Where \( \cos(w, w') \) is the cosine similarity between the words \( w, w' \), and the matrix \( M_{y} \) represents the indicator matrix of the mapping relationship between the words and the original meaning. When \( M_{y} = 1 \), it means that the first word \( w \) has the \( j \)th original meaning. If \( M_{y} = 0 \), it means that the original meaning is not belongs to the word \( w \).

Because the vocabulary is large, the meaning of dissimilar words is actually equivalent to noise for the word, so the ranking according to similarity is set.

For serial number 1, the weight coefficient \( c^{r_i} \) gradually decreases, and the weight coefficient is calculated through the weight \( c \in (0, 1) \) and the drop index \( r_i \).

When the similarity ranking number between the word and the target word is larger (that is, the larger the \( 1 \) is, the less similar it is) and the larger the \( r_i \) is, the weight coefficient is \( c^{r_i} \) more smaller, so that the meaning of more similar words gets more weight, making the prediction result more accurate.

The above-mentioned semantic primitive prediction methods such as SPWE and SPSE only use the word vector, that is, the external context information of the word. The final semantic primitive prediction completely depends on the quality of the word vector. Therefore, when the word has no word vector, that is, the case of OOV words, and low frequency in the case of words, the above methods are either unpredictable or have poor results. In order to solve these problems, the literature proposes a method to predict the meaning of the original words combined with the internal character information. A lot of research has been done on word-level natural language processing methods in academia. The research is based on the principles of morphology. In Chinese, words are composed of one or more characters, and most of the characters have corresponding semantic meanings in the words, and these characters with meaning belong to morphemes. Chinese words can be divided into monomorphic words and compound words, and compound words account for a higher proportion.
Since the meaning of a compound word is closely related to its internal characters, whether it is a compound word or a monomorphic word, its semantic meaning can basically be inferred from the internal characters of the word. According to the principle that the semantics and position information of the words can infer the semantics of the words, the SPWCF method is proposed in the literature, which splits the compound words into three parts: beginning, middle, and ending. It is assumed that words with similar structures have similar meanings. According to the collaborative filtering algorithm, the original prediction result of the new word can be obtained. The prediction probability calculation formula is shown in formula (2).

\[
P(s_{ij}, w) = \sum_{p \in \{B, M, E\}} \sum_{c \in \pi_p} p_p(s_{ij}, c)
\]

\(s_{ij}\) means that for a given word C and a position p, \(\pi_p\) can be \(\pi_B, \pi_M, \pi_E\), representing the beginning, middle and end of the word respectively, where \(p_p(s_{ij}, c)\) in the above formula is:

In this document, the SPCSE method for predicting semantic origin using the embedding of words and semantic origin is also proposed. This method follows the principle of matrix decomposition of SPSE method, and simultaneously carries out the calculation of the matrix of "word-Yuanyuan" and "Yuanyuan" matrix. The difference of matrix decomposition is that the word vector used is a pre-trained word vector to use the information of the words inside the word. In the CSP method, including the SPWE method and SPWCF, only the relevant information of the word itself is used, and neither the word vector nor the words inside the word use external knowledge. As a semantic description, the meaning of the original is similar to the definition. The definition is a very direct and supervised semantic information of a word. If dictionary definitions are used to make Liyuan prediction, external knowledge is introduced, which will help the accuracy of Liyuan prediction. This method is based on the Seq to seq model, which is a language model widely used in machine translation. It can make predictions from one sequence to another. Usually, the two sequences are in two different languages. The characteristic of the sequence is that there is an order dependency between the preceding and following units, and the following units depend on the previous unit. The LD-Seq to seq method regards the original meaning of a word as a sequence with a weak order relationship, thus applying the Seq to seq model, as shown in the figure below.
In this method, in order to find the optimal label sequence (the original sequence of the target word), it needs to be optimized through training. Conditional probability \( p(s|D) \).

\[
p(s|D) = \prod_{i=1}^{k} P(s_i|s_1, s_2, \ldots, s_{i-1}, D)
\]

Where \( D=(d_1, d_2, \ldots, d_i) \) is the embedding of the word in the definition, this method does not use the word after word segmentation, the number of words in the sentence sequence is 1, \( s' \) is the set of meanings, and the length is \( k \). The next yuan \( s_i \) in the prediction result is affected by the prediction results of the previous \( i-1 \) yuan. But in fact, there is no clear order relationship between Liyuan. In order to solve the problem of the order of the predicted semantic sequence, the loss function in the LD-Seq to seq method uses a custom soft loss function instead of the hard-cross entropy loss function that is usually used in prediction or classification tasks. To alleviate this problem to a certain extent, it is called the Label Distributed Seq to seq model [3]. The definition data of this method adopts the encyclopedia description and dictionary definition given in Baidu Encyclopedia and Baidu Dictionary, and it is the first method to combine word definitions to do the original prediction. However, this method regards the original set of meanings as an ordered sequence, the modeling method is not suitable, and the input of the model only uses the word-level information in the definition, and does not fully dig out the richer semantic information in the definition.

4. Intelligent natural language processing auxiliary platform system design

The knowledge-based recommendation system can be said to be a kind of reasoning technology to a certain extent. It does not depend on the user's preference itself, but sets specific rules in a specific application field for case-based reasoning. This recommendation method mainly lies in the difference of the functional knowledge used. The functional knowledge is the relationship between the user's interpretation needs and the recommendation, and it is related to the knowledge of how a specific project meets a specific user. As far as the knowledge recommendation of the intelligent natural language processing auxiliary system of university libraries is concerned, it is a recommendation result library that requires knowledge, and students' registration information, score information, course information, etc. can all be knowledge structures that support reasoning. These knowledge structures are different from pure user preferences. Therefore, the natural language processing auxiliary system in this article first needs to obtain the user knowledge and library collection knowledge of college students and generate a basic recommendation list to users through functional knowledge reasoning or language matching.

4.1. Agent architecture selection

The construction of an intelligent natural language processing auxiliary platform for university libraries is a complex project that requires multi-agent collaboration, which involves system structure issues. There are three common architectures for multi-agent collaboration: network structure, hierarchical structure, and alliance Structure [3]. These three common multi-agent architectures have their own characteristics. Any two Agents in the multi-agent system structure of the network structure are in a peer-to-peer relationship [4]. The advantage is that convenient direct communication can be established. The disadvantage is that a large amount of communication information will be generated and stored, which will lead to the efficiency of the entire system for a complex system. low. The hierarchical structure greatly reduces the amount of communication information storage, but even the same level of communication must be completed through the instructions of the upper layer, resulting in a decrease in communication efficiency. The multi-agent of the alliance structure can divide the agents with similar attributes and application functions together to form an agent alliance, and each alliance has a management agent. In this way, the appropriate communication method can be selected
4.2. System process

According to the demand analysis and the selection of the Agent system structure, the application process of the intelligent natural language processing auxiliary platform of the university library based on the Agent constructed in this paper is shown in Figure 3.

According to the system flow chart, when a specific college student is using the intelligent teaching assistant system, it is first accessed through the unified single sign-on authentication system of the university campus network, and then the corresponding Agent will obtain the screen resolution of the device used by the student. Information, so as to adapt to the display of the login device. Then, the teaching assistant platform further obtains the student’s relevant registration information, grade information, course information, course progress information and other basic information. If the student logs in for the first time, he needs to access the recommendation engine in the department alliance first, if he has already accessed Then it directly obtains the recommended information that the library continues to mine and perform basic classification for each department. Further according to the student's browsing behavior, activity behavior and other behaviors that the system has been able to obtain to determine the interest preference of a specific reader. Finally, the display recommendation results are generated under the comprehensive recommendation of the three results of the library basic mining classification information, the recommendation information in the college and department alliance, and the reader's personal preference. The user can also revise this comprehensive recommendation result. The number of revisions reaches a certain number. The system will judge that the student’s main learning tendency has the possibility of interdisciplinary, and it will be connected to the Agent alliance of other departments to obtain Cross-faculty, cross-disciplinary recommendation. As long as the student does not shut down the operating system, the system will continue to monitor various related behavioral information of the user, correct the information, modify the recommended content, and refresh it regularly.

4.3. System Architecture

The system architecture is based on the classic B/S architecture foundation, which is conducive to cross-platform browsing. At the same time, an intelligent agent layer is added to form a new B/S/A architecture form, as shown in Figure 4.
The Browser layer is mainly used to manage the user interface, realize the interaction of information input and output, and automatic collection. The Server layer is mainly composed of a cluster of database servers, Web servers and recommendation servers, which completes database synchronization, connection, Web interaction, recommendation data storage, exchange and implementation. Each library can choose a suitable solution according to the specific conditions of data processing and storage. The Agent layer realizes the user's intelligent recommendation. The layer can contain a variety of intelligent agents, and it is also very convenient to expand and access various intelligent agents. For example: Interface Agent is used to determine which device the user uses to log in to the system, and automatically adapt to the user interface with different resolutions. Monitoring Agent is mainly used to monitor the behavior of their respective users. In a word, the user interface agent, monitoring agent, recommendation engine agent, data mining agent, satisfaction agent, preprocessing agent, etc. have formed an intelligent natural language processing auxiliary platform for university libraries [5].
The recommendation engine Agent is the core component of the entire Agent layer. The internal structure shown in Figure 5 shows that the agent itself is not an independent individual, it is composed of multiple sub-agents. The figure shows the data mining agent, recommendation control agent, and recommendation production agent, which correspond to the books of the data service layer. Library database, user interest model library, recommendation algorithm library and recommendation record library

![Figure 5. Recommendation engine Agent framework diagram](image)

5. Related Discussion
At present, most of the NLP researchers in the world have a computer science background, and have largely ignored the linguistic theory that should be the cornerstone of NLP, and the formal analysis of language has fallen into a low ebb. In view of this, the author will discuss the interdisciplinary status and historical role of language forms.

5.1. Interdisciplinary status of NLP formal analysis
The development of computational science and emerging cognitive science has also inspired formal research on language. For example, the discovery of neural networks and the application of artificial neural networks have shifted the focus of NLP from linear grammar rules to random probabilities in network structures. The complex network is not only considered as an ideal model for semantic formalization, but its huge three-dimensional space can also allow grammatical rules to emerge naturally and be represented in the system in a dynamic, probabilistic, and parallel processing manner. Under this understanding, NLP

There have been various unification theories and algorithm techniques combining probability [6].

5.2. The historical role of NLP formal analysis
The development of NLP formal analysis theory and methods seems to have lagged behind other computing fields. For example, formal models based on knowledge evolution systems are still relatively rare in NLP, and evolution is precisely an important content of language as a system development and individual ability development. In addition, there are CYK parallel algorithm, ATN procedural algorithm, and sub-sentence theory in the algorithm technology and theory based on NLP formal rules, which have not been fully utilized under the new technical conditions. Therefore, there is still a lot of room for development in NLP formal analysis.
5.3. Development prospects of NLP formal analysis theory

The successful application of neural networks and deep learning in NLP is not only a product of rationalism, but also a result of empiricism. On the one hand, neural networks are the result of the scientific development of biology and complex systems. The application of NLP in NLP is a top-down deduction of a general principle, and a systematic and comprehensive theoretical explanation needs to be obtained in the formal research of natural language. In particular, the various matrix dimensionality reduction techniques used in the deep learning process have not yet received a better mathematical explanation. On the other hand, the application of neural network models and deep learning algorithms in the NLP field needs to rely on big data. This kind of statistical research is an empirical induction method, which is close to the facts but far from the scientific theory. There is a lack of linguistic explanation in NLP. The philosophical basis of formal analysis theory is rationalism. The essence of this paradigm that has dominated scientific research for a long time is to seek the solution of the problem and seek the simplest expression. The goal of rationalism in the field of artificial intelligence is to find a formal structure that converts human intelligence activities into abstract symbolic system operations. In order to find well-posed problems, early rationalists put forward many restrictions in NLP, such as atomic units, linear order, finite sequence, etc., in order to limit the processing objects and processing processes within a controllable range. With the introduction of probability ideas and the use of sub-symbol continuous variables, semioticist overcomes the strong tendency of early atomism.

6. Conclusion

The improvement of computing power and the application of deep learning are not enough to deny the role and value of natural language formal analysis in the history of NLP and computational science. There is still room for improvement and development in the theories and methods of language formal analysis. A comprehensive, systematic, and profound understanding of various formal analysis theories of natural language is still very important for the further development of NLP and cognitive science. NLP technicians can extract ideas with theoretical value in the practice of various algorithm improvements. Linguists and language natural language processing workers should also constantly update their knowledge, pay attention to measurement methods, actively use cutting-edge technologies, and jointly promote computing science. And the development of cognitive science.

References

[1] Pons, E., Braun, L. M., Hunink, M. M., & Kors, J. A. Natural language processing in radiology: a systematic review. Radiology, vol. 279, pp. 329-343, February 2016.
[2] Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. Clinical natural language processing in languages other than english: opportunities and challenges. Journal of biomedical semantics, vol. 9, pp. 12, January 2018.
[3] Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., & Xu, H. CLAMP—a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association, vol. 25, pp. 331-336, March 2018.
[4] Young, T., Hazarika, D., Poria, S., & Cambria, E. Recent trends in deep learning based natural language processing. IEEE Computational intelligenCe magazine, vol. 13, pp. 55-75, March 2018.
[5] Yim, W. W., Yetisgen, M., Harris, W. P., & Kwan, S. W. Natural language processing in oncology: a review. JAMA oncology, vol. 2, pp. 797-804, June 2016.
[6] Castro, V. M., Dligach, D., Finan, S., Yu, S., Can, A., Abd-El-Barr, M., ... & Savova, G. Large-scale identification of patients with cerebral aneurysms using natural language processing. Neurology, vol. 88, pp. 164-168, February 2017.