ISSN: 2664-9640

Word similarity calculation method based on natural language processing context

Jiayang Dai, Dong Zhou*

Hunan University of Science and Technology &School of Computer Science and Engineering, Hunan, China

Abstract

In recent years, with the rapid development of information technology in China, in order to improve the accuracy and efficiency of daily language processing in natural language processing context system, information technology personnel have been constantly innovating the computing method of context similarity processing. In the process of natural language processing, the language system analyzes and studies the context information, and checks the word similarity through information technology and function related content, so as to avoid word repetition and other phenomena. This paper mainly analyzes the word similarity calculation background, method and experimental results. The experimental results show that the similarity calculation method of context words in natural language processing can effectively distinguish word repetition probability in the system and screen out repeated words and even paragraphs.

Keywords

Natural language processing context; Word similarity; Calculation method.

1. Background of word similarity calculation in natural language processing context

Word similarity mainly refers to the words or contents with high frequency appearing in the whole article or paragraph, which are identical with other articles. The natural language processing context system can screen out these repeated words and paragraphs, so that users can understand them. At present, the translation software, information retrieval has been widely used this kind of natural language processing technology, in use process, not only for high repetition rate of filtered words, paragraphs, can also use these content tagging, effectively enhance the accuracy in natural language processing technology of context and screening rate, greatly improve the efficiency of the user. In addition, in the application of NLP context technology, words with high frequency can be classified, and these words can be statistically summarized to improve the application rate of natural language. Language similarity calculation is not limited to the functions of duplicate checking on websites such as CNKI. Users can also search for relevant literatures through CNKI and enrich their own manuscripts by referring to the contents of relevant literatures, so as to effectively improve the acceptance rate of manuscripts.[1]

2. Computational method of word similarity in natural language processing context

2.1. Word similarity calculation method based on semantic dictionary

Semantic dictionary word similarity calculation method is a simple, based on the language class, and a method of artificial intelligence technology, the technology is mainly from the semantic dictionary, the word concept transposition, forming a synonymous with the relationship

ISSN: 2664-9640

between up and down, at the same time the calculation way to build a relationship tree, through the way of relationship tree network diagram, making it easy for users to access. This calculation method is relatively simple and easy for users to operate and understand, but the content of the screening is not very accurate, prone to subjective influence, unable to express the content of words intuitively and objectively. At present, English dictionaries such as Word Net and Frame Net are used for reference for this type of calculation, while Common Chinese dictionaries include Chinese Concept Dictionary and Thesaurus.[2.3.4]

2.2. Word similarity calculation based on statistics

By summarizing and sorting out frequently searched words, the computer engineers build their own language system and build a large-scale language library. Natural language processing technology in the context context similarity on in-depth research and analysis, natural language processing can be found on the vector space model is natural language processing technology context is one of the main analysis model, the space model, complex algorithm can be divided, a set of different meanings of key at the same time, categorizing these key words, Thus, word similarity can be screened and classified effectively. In addition, users can also filter the models in the vector space through context retrieval, and calculate the distance between the words in the vector space model, so as to further realize data informatization and automation. Finally, technicians store the words so they can be consulted later.[5.6]

2.3. Compare the similarity between semantic words and corpus words

Technical personnel in order to better find convenient for user operation context of natural language processing technology, based on semantic similarity of words and word corpus, this paper compares and analyzes the method of this article mainly from the two sets of technology, set up conditions, basic conditions, basic theory, advantages and disadvantages as well as the evaluation method for simple comparison and analysis, specific content see table 1.

Table 1 compares the similarity between semantic words and corpus words

	Word similarity calculation in semantic dictionary	Word similarity calculation in corpus
methodology	Rationalist methodology	Empirical methodology
Conditions for the validity of the method	There is semantic correlation between the two groups of words, and there is a bridge in the conceptual structure hierarchical network diagram of the two groups	Word semantics can be analyzed in combination with paragraph context, while word semantics only exist in paragraph similar environment
Basic conditions	Semantic dictionary	Large scale corpus
The main theoretical basis	Tree diagram	Vector space
advantages	Simple, direct, literal difference, statistical similarity of words	It can objectively reflect the voice, syntax, semantics and other characteristics of words, and can recognize many characters or characters that cannot be recognized manually
disadvantages	Easy to be affected by the external environment, can not	Data is easy to miss, noise interference and so on

ISSN: 2664-9640

	objectively reflect the meaning of words	
Evaluation method	At present, there is no unified evaluation method	Corpus for unified adjudication

3. Analysis of experimental results of word similarity in natural language processing context

This paper uses natural language processing (NLP) context technology to retrieve word similarity. In order to better evaluate NLP, this paper analyzes and studies the effect of NLP context technology through experiments. When users use natural language processing context technology, they will directly, accurately and unerringly classify these words. The author of this paper will screen and sort out the words with high similarity to the "generated" words, as shown in Table 2 below.[7.8]

Table 2 Other words that are more similar to "produced" words

words	Semantic similarity
appear	85.46%
happen	80.79%
achieve	76.99%
The formation of	73.56%
create	70.81%
cause	68.74%
Put forward	66.38%
cause	48.46%

As shown in Table 2 above, it can be seen that there are many words with similar meanings to the word "produce", which can truly and accurately reflect the similarity of the word.

4. Conclusion

To sum up, it can be seen that natural language processing context technologies include automatic retrieval, text classification, automatic solution, word translation and so on, which can effectively assist users to detect word similarity and reflect word similarity in a real and objective way. In addition, natural language processing context technology combined with Internet information technology to sort out and summarize these words with high similarity, so as to effectively improve the efficiency and accuracy of word screening. However, due to the problem of data loss, NLP still needs to be perfected by technicians.[9]

References

[1] Sa A , Aa A , Rmrg A . Detecting Semantic Similarity Of Documents Using Natural Language Processing[J]. Procedia Computer Science, 2021, 189:128-135.

[2] Zhang P , Huang X , Wang Y , et al. Semantic Similarity Computing Model Based on Multi Model Fine-Grained Nonlinear Fusion[J]. IEEE Access, 2021, PP(99):1-1.

[3] Guan X , Han J , Liu Z , et al. Sentence Similarity Algorithm Based on Fused Bi-Channel Dependency Matching Feature[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2019, 34(5).

[4] Wang Y . Similarity detection of English text and teaching evaluation based on improved TCUSS clustering algorithm[J]. Journal of Intelligent and Fuzzy Systems, 2020, 40(4):1-11.

ISSN: 2664-9640

[5] Navigli R , Martelli F . An overview of word and sense similarity[J]. Natural Language Engineering, 2019:1-22.
[6] Wang Bin. Research on automatic alignment of Chinese-English bilingual corpus [D]. Beijing: Institute of Computing Technology, Chinese Academy of Sciences,1999. (in Chinese)
[7]Jijian Xie ,Chengping LIU . Fuzzy Mathematics Method and Its Application [M]. Huazhong University of Science and Technology Press, 2006.15-37.
[8] Yu Chao. Research and application of word similarity calculation based on KNOnet [D]. Shenyang: Shenyang Institute of Aeronautical Technology,2006. (in Chinese)
[9] Guo Li. Word similarity calculation based on context and its application [D]. Shenyang: Shenyang Institute of Aeronautical Technology,2009. (in Chinese)