Using Two Formal Strategies to Eliminate Ambiguity in Poetry Text

Wei Hua, Shunpeng Zou, Xiaohui Zou, Guangzhong Liu

Using Two Formal Strategies to Eliminate Ambiguity in Poetry Text

Wei.Hua[1],Shunpeng.Zou[2], Xiaohui.Zou [3][0000-0002-5577-8245] and Guangzhong.Liu[4]

[1] College of Information Engineering ShangHai Maritime University, ShangHai, China
201740310003@stu.shmtu.edu.cn
gzhliu@shmtu.edu.cn
[2] China University of Geosciences (Beijing), Beijing, China
407167479@qq.com
[3] Sino-American Saerle Research Center, Beijing, China
949309225@qq.com

Abstract. The purpose of this paper is to compare the two major types of formalization strategies through the disambiguation of natural language textual ambiguities. The method is: The first step is to select the same text. Using poetry as an example, two types of formal strategies are used to resolve the ambiguities that exist. The second step is to analyze the limitations of the first formal path, at the same time, using traditional artificial intelligence methods and a new generation of artificial intelligence. The third step is to use the double-word board tools and methods to do the same thing. The result is that using the first path, whether based on rules (traditional artificial intelligence methods) or on statistical and machine learning, especially deep learning (a new generation of artificial intelligence methods), only local solutions can be obtained; With the checkerboard tools and methods, the overall solution can be obtained. This shows the unique advantages of the second path. Its significance lies in: using the double-word chessboard tool and method (second path) can solve the common problems faced by traditional artificial intelligence and new generation of artificial intelligence, and how to eliminate the ambiguity of natural language texts. The most important thing is that it has a new role. The most typical is to construct a knowledge base of the subject through the acquisition of knowledge and formal expression of experts, so as to gradually resolve a series of ambiguities between natural language (text) processing and formalized understanding.

Keywords: Two Formal Strategies, Eliminate Ambiguity, Natural Language Textual, Ambiguities Artificial Intelligence Methods

1 Introduction

Ambiguity is a type of meaning uncertainty giving rise to more than one plausible interpretation. It generally exits in our language and expression, being ambiguous is therefore a semantic attribute of a form (a word, an idea, a sentence, even a picture) whose meaning cannot be resolved according to a simple rule or process with a finite number of steps. As the result, semantic disambiguation plays an important role in natural language processing (NLP) and many scholars have been spending tremendous effort on the problem for decades. However, the development of disambiguation technique stagnated in a long term [1] until the significant breakthrough was made on frameworks and algorisms of neural network in last decade [2]. By building computational models based on statistic theory, many significant research results and solutions regarding disambiguation have been obtained rather than by using traditional linguistics. However, the challenge and obstacle still exist, even by using neural network or deep learning, the accuracy in tasks of disambiguation has still a lot space to improve especially in the field of Chinese poetry understanding. The main reason why this task is difficult is that even neural network and deep learning emulate successfully how human's brain works on the task of language processing, the fundamental form of these new methods are still based on Aristotle 's formal logic and Frege' s mathematical logic[3], which means only the form of programming languages is involved into the NLP tasks so that whether rules based on traditional artificial intelligence methods or on statistical methods and machine learning, especially deep learning (a new generation of artificial intelligence methods), only local solutions can be obtained. For convenience to describe the path, here we define this form as "first type of formal strategy". .

In our research, the target is to find a new path to break through the bottleneck that "first type of formal strategy" has to face and then to verify its effectiveness. In this paper we present a way to the second path – "The second type of formal strategy" [3], to resolve the problem. According to the idea of second path, overall solution can be obtained with the checkerboard tools and methods .The tool based on second path combines Chinese characters and English, binary and decimal systems, decimal and Chinese characters into "double-words" chessboards. Due to the difference between language structures, the task of semantic disambiguation of Chinese poetry is more difficult than that of modern Chinese, in the meantime, being lack of sufficient corpus leads to the limitation of statistics-based disambiguation. Through comparison of the experimental results, we provide references for further study of two formal strategies.

2 Related Work

To verify the effects of two formal strategies, we design a set of experiments for each type and build a system based on the second formal strategy. We call this system "Double-word board" [4][5]. The same Chinese poem text is fed to both of the systems to test the effect of disambiguation, and finally we make the comparison of the experimental results.

2.1 Build a double-words chessboard and Chinese language chessboard spectrum

The tool based on “The second natural language formalization strategy” is called "**Double-Word Board**" [5]. It contains two main components: the Chinese character board encoded with digitals and the Chinese character board. “Double-word board” is the linkage function between digital and textual of conjugate matrices, binary and decimal codes and English and Chinese and its alternative bilingual. We can think the board as an expert knowledge acquisition system that machine cooperatively builds with human. Through the human-machine interface, Chinese characters in the poem are marked one by one with digital codes. Encoded matrix/table plays a key role in disambiguating the poem text, its digital codes and number combinations indicate the relationship of Chinese characters in context, and the chessboard stores the codes as reusable rules.

2.2 Experiments

We design two test cases for each experiment to verify our language chess-board(Double-Word Board) and its application are effective and accurate in the task of disambiguation in poetry text. We use a Chinese poem as the input of test case 1 (**Fig**.**1**):
“床前***明月***光，疑是地上霜。举头望***明月***，低头思故乡。”

Fig. 1. Formal chessboard(double-word board) in Test case1.

In this Chinese poem, the phrase “明月” appears in the first and the third sentences repeatedly with different interpretations. The phrase appears in the first sentence means “moonlight” and the different meaning, “a bright moon” appears in the third sentence.

We use the built-in encoding system to create codes for each character in this poem and then we can obtain a table containing mapping relations between code and Chinese characters. (**Table 1.**)

Code	Character sample
…	…
564	已
565	*明*
589	打
594	*月*
…	…

Table 1. Create codes for each character in the poem

Encoding system is not complex. We pick out the polysemic phrase “明月” in this poem to demonstrate how the chessboard (double-word board) works. Human can distinguish the different interpretation of this word in context positions. As our indication, number "565" represents "明", and “594” represents “月” are record. Therefore, "明月" can be combined into a new code composed of two numbers, as the result, information containing character sequences is also recorded according the context of the poem sentences.

Next step, in the similar way, we create codes for three-character combination and this set of code are stored in chessboard. Disambiguation of man-machine interaction is needed only once for each phrase.

Our goal is to make comparison between first type of strategy and second type to prove the effectiveness in these two strategies. Thus, we setup test case 2 , inputting the same poem text into a statistics-based computational model so that we can observe the effect and the difference of disambiguation between by using deep learning or machine learning [6] and by “double-word board”. Statistics-based software system(deep learning or machine learning system) mostly needs a dataset that its function is quite similar to the use of code in double-word chessboard. But the nature of datasets is different with that of codes in chessboard radically. Datasets do not indicate the ambiguity clearly, on the contrary, codes stored in chessboard can functionally eliminate ambiguities in the poetry text as the code based on human’s understanding and the knowledge is transferred to the system.

2.3 Public datasets

We found some public corpus containing the poems we need in our experiment. “Wiki corpus”[7]、 “Literature 100”[8] are the two widely used poetry corpus.

Two word string	Word frequency
…	…
日月	508
…	…
月明	515
…	…
天地	586
…	…
明月	896
…	…

Table 2. shows that the system can only counts the number of appearance of “明月”(which means “bright moon or moonlight”), it cannot accurately indicate the different meanings of phrase“明月” in different context positions.

Then we give the second experiment. We pick another Chinese poem as the input text sample:

慈母手中线，游子身上衣。临行密密缝，意恐迟迟归。谁言寸草心，报得三春晖。

Suppose we do not know this poem before. In fact, when someone is learning a foreign language, it is quite often to read obscure sentences those are hard to understand. The situation gets worse in Chinese poetry. The main reason is the fact that non-native Chinese speakers are hard to segment Chinese sentences into phrases correctly meanwhile Chinese sentences usually omit sentence elements unregularly. At present, word segmentation based on statistical techniques has been well developed [9]. There are many open-sourced components available for building a natural language processing system and the performance is satisfied. The commonly used Chinese segmentation methods [10] can fall into three categories: Segmentation methods based on string matching, word segmentation methods based on comprehension, and word segmentation methods based on statistics.

Segmentation method based on string matching: also known as the mechanical word segmentation method, it is a certain or a sort of strategies to look up the Chinese character string in a "full-sized" dictionary, if a word is found in the dictionary then the word segmentation is successful.

The word segmentation method based on comprehension: This method is to design a sort of algorisms to make a computer emulate the process of a person's understanding a sentence. The basic idea of this method is to syntactically and semantically analyze the sentence segmentation, so as to deal with ambiguity with syntactic and semantic information. It usually contains three parts: word segmentation subsystem, syntax and semantics subsystem and general control. Under the coordination of the general control, the word segmentation subsystem can obtain syntactic and semantic information about words, sentences, etc. to judge the ambiguity, that is, it simulates the process of human understanding of the sentence. This segmentation method re-

quires a lot of linguistic knowledge and information. Because of the generality and complexity of Chinese, it is difficult to make Chinese language information into a form unified that can be directly read by machines. Therefore, word segmentation system based on understanding is still in the research stage.

The statistical word segmentation method is that sentence segmentation can be applied on unknown texts with rules learned from a huge amount of corpus segmented with statistical machine learning techniques. For example, the maximum probability word segmentation method and the maximum entropy word segmentation method are commonly used in the task of word segmentation. As the establishment of large-scale corpus, the research and development of statistical machine learning methods, statistic based methods have gradually become the mainstream in the field of Chinese word segmentation.

Here we list some mainstream models: N-gram, Hidden Markov Model (HMM), Maximum Entropy Model (ME), Conditional Random Fields (CRF[11]) etc.

In practical applications, the word segmentation system based on statistics needs a dictionary to perform string matching. Meanwhile, statistical methods combine string frequency with string matching, so as to make word segmentation perform faster , efficiently with function of recognition of new words and automatic elimination of ambiguity.

However, at current stage, these three methods perform still not well enough in the task of word segmentation in Chinese poetry text due to the lack of word dictionaries customized for Chinese poetry.

To this problem, we try the word segmentation with double-word board to see if we can make any improvement on the task.

The board disorder the characters in the poem and we can obtain a list of two-to-five characters combinations:

身，心，意，行，缝，恐，归，言，报，母，子，手，线，衣，谁，春，晖

慈母，游子，临行，意恐，谁言，报得，春晖，寸草，三春，迟迟，密密

手中线，身上衣，密密缝，迟迟归，寸草心，三春晖

慈母手中线，游子身上衣，临行密密缝，意恐迟迟归，谁言寸草心，报得三春晖

Similarly, according to the codes of single Chinese character, two-to-five word combinations are stored into the chessboard that means we transfer our knowledge to the system. After man-machine cooperation, as the persons who never read this poem we can roughly understand the correct word segmentation in the poem. Only corpora segmented correctly has the value for further processing.

In this experiment, we use the statistics-based segmentation tool jieba [12]to perform segmentation for the poem text and the sentences segmented list is:

慈|母|手中|线|游|子|身|上衣
临|行|密|密|缝|意|恐|迟|迟归
谁|言|寸|草心|报|得|三春|晖

As we can see a few segmentation mistakes still made in segmentation.

3 Discussion

Eliminating ambiguity is an important function in Natural Language Processing (NLP). This paper compares the disambiguation effects of two formal strategies by using the same poems for different scenarios of ambiguity through two sets of experiments. We can observe it has a better performance in task of disambiguation with bilingual chessboard tools. The compared results show we can still find out the relationship between the two strategies and their respective advantages and disadvantages.

Through the experiments we can see that statistics still do not work well in task of disambiguation in poetry texts. The reason for this situation is still due to the influence of the ambiguity that generally exists in languages. When dealing with ambiguity, statistical algorisms still have no enough ability to obtain global solution like human does.

Therefore, the bilingual chessboard used in this paper, that is, the combination of the formalization of the two formal strategies[13], allows the system to obtain a global optimal solution through human-computer interaction and effectively solves the problem of ambiguity in Chinese poetry. Combining the advantages of programming language and statistical techniques, the knowledge base is constructed to eliminate ambiguity in the language.

4 Future work

Text functions are recorded in the form of Chinese or English words, and the order and position of the lattice in a particular board or matrix are also relatively constant. Although their combination can be ever-changing, but in the linkage function of the constraints, there are still laws following the rules. This is the role, value and significance of the three types of identities and their corresponding analytic geometric representations. Among them, the three types of identities embody the basic laws of three kinds of information, and the corresponding analytic geometric expression can be presented through the twin chessboard (double-chess board) and play a role in the process of man-machine collaboration, which is expressed as an expert knowledge Acquisition. Demonstration of “The second type of formal strategy” is still a rudiment in our paper. Our further work will be still working on the novel idea of knowledge transformation to computational system effective and efficient.

5 Acknowledge

All sources of funding of the study have been disclosed.

References

1. D. L. Waltz.: ON UNDERSTANDING POETRY, Theoretical Issues in Natural Language Processing (1975)
2. Ming-Hong Bai., Keh-Jiann Chen.,Jason S. Chang.: 利用雙語學術名詞庫抽取中英字詞互譯及詞義解歧 (Sense Extraction and Disambiguation for Chinese Words from Bilingual Terminology Bank) [In Chinese]. Proceedings of the 17th Conference on Computational Linguistics and Speech Processing，pp. 305-316, September (2005)
3. XiaoHui.Zou., Shunpeng ZOU., Lijun KE.: Fundamental Law of Information: Proved by Both Numbers and Characters in Conjugate Matrices. IS4SI (2017)
4. XiaoHui.Zou.: Bilingual information processing method and principle. (2015)
5. Shunpeng Zou.: "FORMAL BILINGUAL CHESSBOARD SPECTRUM：SHOW THE OVERLAPPING BETWEEN LANGUAGE AND MIND",AAAS 2017 Annual meeting, February (2017)
6. Chao-Lin.Liu., Chun-Ning.Chang.,Chu-Ting.Hsu.,Wen-Hui Cheng.,Hongsu Wang .,Wei-Yun Chiu.:《全唐詩》的分析、探勘與應用－風格、對仗、社會網路與對聯(Textual Analysis of Complete Tang Poems for Discoveries and Applications - Style, Antitheses, Social Networks, and Couplets)[In Chinese]. Proceedings of the 27th Conference on Computational Linguistics and Speech Processing (ROCLING 2015),pp. 43-57 (2015)
7. https://zh.wikisource.org/zh-hant/全唐詩
8. http://www.wenxue100.com/
9. Manex.Agirrezabal., Iñaki.Alegria., Mans Hulden.: Machine Learning for Metrical Analysis of English Poetry. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers pp.772-781, December (2016)
10. Wanying.Jin.: CHINESE SEGMENTATION DISAMBIGUATION, COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics (1994)
11. John Lafferty., Andrew McCallum.,Fernando C.N. Pereira.: Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data. Department of Computer & Information Science. (2001)
12. https://github.com/fxsjy/jieba
13. Shunpeng.Zou.,Xiaohui.Zou.: Understanding: How to Resolve Ambiguity. International Conference on Intelligence Science pp.333.(2017)