Hindawi
Complexity
Volume 2024, Article ID 9782634, 1 page
https://doi.org/10.1155/2024/9782634

WILEY | Hindawi

Retraction

Retracted: Numerical Simulation of Ambiguity Resolution in Multiple Information Streams Based on Network Machine Translation

Complexity

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] L. Wang and Q. Ai, "Numerical Simulation of Ambiguity Resolution in Multiple Information Streams Based on Network Machine Translation," *Complexity*, vol. 2020, Article ID 7278085, 10 pages, 2020.

Hindawi
Complexity
Volume 2020, Article ID 7278085, 10 pages
https://doi.org/10.1155/2020/7278085

Research Article

Numerical Simulation of Ambiguity Resolution in Multiple Information Streams Based on Network Machine Translation

Lei Wang[1] **and Qun Ai**[2]

[1]*Department of Foreign Language, Jilin Business and Technology College, Jilin, Changchun 130000, China*
[2]*Department of Basic Education, Jilin University, Jilin, Changchun 130000, China*

Correspondence should be addressed to Qun Ai; sabirin9917@mails.jlu.edu.cn

Received 6 April 2020; Revised 22 July 2020; Accepted 3 August 2020; Published 17 August 2020

Guest Editor: Zhihan Lv

In natural language, the phenomenon of polysemy is widespread, which makes it very difficult for machines to process natural language. Word sense disambiguation is a key issue in the field of natural language processing. This paper introduces the more common statistical learning methods used in the field of word sense disambiguation. Using the naive Bayesian machine learning method and the feature vector set extracted and constructed by the Dice coefficient method, a semantic word disambiguation model based on semantics is realized. The results of comparative experiments show that the proposed method is better compared with known systems. This paper proposes a method for disambiguation of word segmentation in professional fields based on unsupervised learning. This method does not rely on professional domain knowledge and training corpus and only uses the frequency, mutual information, and boundary entropy information of the string in the test corpus to solve the problem of word segmentation ambiguity. The experimental results show that these three evaluation standards can solve the problem of word segmentation ambiguity in professional fields and improve the effect of word segmentation. Among them, the segmentation result using mutual information is the best, and the performance is stable.

1. Introduction

Word sense disambiguation is to determine the most exact semantics of a polymorphic word according to its context and locale [1]. Chinese is the most spoken language in the world, and it is difficult to avoid the problem of polysemy. According to statistics, ambiguous words account for about 14% of the total vocabulary of the Chinese dictionary, and these ambiguous words are often commonly used words [2]. Statistics on the authoritative Chinese disambiguation corpus show that these ambiguous words are used very frequently, about 42% [3]. The universality of ambiguous word distribution makes word sense disambiguation an important link in many applications related to natural language processing, such as machine translation, information extraction, and content analysis [4, 5].

The earliest Chinese word segmentation method was a word segmentation method based on "lookup dictionary" [6]. The idea of this method is to read the entire Chinese sentence and then mark all the words in the dictionary separately. When encountering a compound word (such as Peking University), the longest word match is found. The string is split into individual words. This word segmentation method is not efficient, but its proposal lays the foundation for Chinese automatic word segmentation technology [7]. Relevant scholars have theorized the Chinese word segmentation method and proposed the "minimum number of words" segmentation theory; that is, each sentence should be segmented with the least number of words [8]. This word segmentation method is an improvement on the "word dictionary" word segmentation method, which has promoted the development of Chinese word segmentation technology. Researchers believe that, for computers to reach the level of natural language processing by humans, computers must be able to automatically disambiguate ambiguous words in a specific context and choose the most precise

meaning [9–11]. Although the word sense disambiguation system was only a subsystem in the machine translation system at that time, the context window and semantic consistency proposed by it were still the basis for the current research on word sense disambiguation. Among them, the context in which ambiguous words are located is an important condition that affects the accuracy of word sense disambiguation [12]. After that, with increasing attention to word sense disambiguation, more and more experts and scholars have proposed more solutions to word sense disambiguation [13, 14]. With the deepening of research, due to the lack of relevant resources and calculation methods at that time, scholars realized that word sense disambiguation was a very complicated problem and it was difficult to overcome [15]. Bar-Hillel pointed out that word sense disambiguation was impossible to achieve with the technical conditions at the time and was theoretically not feasible [16]. The method of automatically expanding WordNet uses a lot of semantic relationships from Wikipedia resources to expand WordNet. First, it establishes a mapping between Wikipedia pages and WordNet and then converts the pages into WordNet. Test results show that this method improves baseline and can use more disambiguation information to achieve higher disambiguation goals [17]. With the continuous updating of technology and continuous improvement of machine performance, techniques such as machine learning and corpus are used in word sense disambiguation. During this period, supervised, semisupervised, and unsupervised word sense disambiguation techniques have been fully developed. Relevant scholars have proposed a completely unsupervised method based on the topic word frequency estimation model [18]. This method can be applied to any part of the speech manuscript without the need for a hierarchical corpus or parallel text. The library is highly portable. Furthermore, the effectiveness of the method on the main semantic learning and semantic distribution acquisition tasks is proved. Relevant scholars have studied a new and effective fuzzy classification system and applied this system to word sense disambiguation [19, 20]. The system iteratively adjusts the weights of fuzzy rules and adjusts the classifier by adding weights to the rules [21]. Compared with other classification systems, this classification system has achieved good results. The advantage of the unsupervised word segmentation method is that it does not rely on dictionaries and training corpus and does not require training. It can be used for unregistered word discovery without word formation rules. The disadvantage is that this kind of method cannot find low-frequency words, the upper limit of the word segmentation effect is about 0.85, and the word segmentation effect cannot meet the practical application. In practical applications, the unsupervised word segmentation method is generally not used alone. This type of word segmentation method can be used for common word discovery, new words discovery, and other issues and assist in improving the word segmentation effect of existing dictionaries and training corpus word segmentation methods.

For word sense disambiguation, these classification models cannot be simply used, and corresponding improvements should be made on the basis of the original models. This paper uses the sliding word window to extract the semantic-related features of words in the word window and constructs a Bayesian word sense disambiguation model based on semantics to perform the word sense disambiguation experiment. The Bayesian word sense disambiguation classifier based on semantic information is constructed using three different vector sets in feature extraction. The new method is used to verify the word sense disambiguation performance. This paper proposes to use unsupervised learning to solve the problem of word segmentation ambiguity in professional fields. The frequency of the strings in the test corpus, the mutual information of the strings, the boundary entropy of the strings, and the boundary entropy of the single words are used as the evaluation criteria to resolve the ambiguity problem. Experiments show that these three evaluation standards can solve the problem of word segmentation ambiguity in professional fields to varying degrees. The rest of this paper is organized as follows. Section 2 studies the semantic-based disambiguation model of machine translation. Section 3 analyzes the method of disambiguation of word segmentation in unsupervised professional fields. Section 4 summarizes the full text and points out future research directions.

2. Semantic-Based Disambiguation Model for Machine Translation

2.1. Machine Translation System

2.1.1. Language Model. The language model treats sentences as strings, and the probability that each word in the sentence appears as a character in the sentence is random. For a given string $w_1, \ldots, w_n$, the probability of its occurrence can be expressed as

$$P(W) = \prod_{k=0}^{n-1} P\left(w_k \mid w_1, w_2, \ldots, w_{k-1}\right). \tag{1}$$

N-grams are generally used in language models to calculate the probability of an entire string. The probability that the Nth word appears in the N-gram model is only related to the first N-2 words; that is,

$$P\left(w_n \mid w_1^{n-1}\right) = P\left(w_n \mid w_{n-N+2}^{n-2}\right). \tag{2}$$

Linguistic studies have shown that the emergence of the current word is strongly dependent on many of the words before it [22]. Language models provide a way to calculate the probability of a string appearing. The disadvantage is that a large-scale corpus is needed to determine the parameters of the model.

2.1.2. Translation Model. In order to construct a translation model for a phrase machine translation system, we first need to calculate the phrase translation probability and dictionary probability. The phrase translation probability indicates the probability that a phrase on the source language side is translated into a phrase on the target language side. To obtain these two probabilities, four operations need to be performed on the parallel aligned bilingual corpus: word

alignment, word score, phrase extraction, and phrase score. A schematic diagram of a network bilingual information processing system for machine translation is shown in Figure 1.

According to the probability score calculation of the translated sentence, the longer the sentence, the smaller the probability. So when translating, decoders for phrase machine translation tend to choose shorter sentences. Therefore, longer target sentences need to be compensated. The length penalty model calculates the number of words in the translation as a penalty value and adds it to the model, which can be expressed as

$$\Pr(e) = \exp(I). \tag{3}$$

In the formula, I represents the number of translated words. The word punishment model can choose the length of the translation.

2.2. Word Sense Disambiguation Method Based on Statistical Learning. The Bayesian method is implemented using probability calculations. It is inferred from the probability that an event has happened in the past. The Bayesian method is applied to the word sense disambiguation problem as follows:

$$P(S_i \mid \text{Context}) = \frac{P(S_i)P(\text{Context} \mid S_i)}{\sum_{j=1}^{n} P(S_j)P(\text{Context} \mid S_j)}, \quad i = 1, 2, 3, \ldots, m. \tag{4}$$

Word sense disambiguation based on Bayesian method is to judge the classification of word meaning based on the size of the posterior probability. Among them, Context is the context in which the ambiguous word w is located. Context is composed of word units on both sides of w, which provide necessary guidance information for the disambiguation process. The ambiguous word w has m semantic categories $S_1, S_2, \ldots, S_m$. In Context, its true semantic category is S_j. $P(X)$ is the probability that X will occur. In the Context, if the probability of the ambiguous word w taking the semantic category S_j is greater than any other semantic category S_i ($j \neq$ i), then the semantic category of the word w should be determined as S_j. The Bayesian method has the following two commonly used classifications in practical applications.

2.2.1. Multivariate Bayes Based on Berle Effort. First, we find the feature vector $F = (t_1, \ldots, t_m)$ through feature selection, where t_i is the label of each feature. For the dichotomy, this method deduplicates the corpus sample d to obtain the label set $X = (x_1, \ldots, x_m)$, where x_i is the categorical variable value, which is 0 or 1 for the dichotomous case. According to the results of m Bernoulli experiments, the disambiguation corpus belongs to the category C:

$$P(t \mid c) = \frac{(M_{t,c} + 1)}{(M_c + 2)}. \tag{5}$$

To prevent the denominator from being zero, we use Laplace's method for smoothing. M_t,c is the number of texts that belong to category c and feature t_i, and M_c is the number of texts that belong to category c.

In order to prevent the result from underflowing, logarithmic processing is generally applied to the probability value in the application. Word sense disambiguation classification criterion is $T = 0$; it judges an ambiguity word classification.

2.2.2. Polynomial Bayes Based on Boolean Attributes. Polynomial Bayesian method based on Boolean attribute is similar to polynomial Bayesian method based on word frequency. The attribute value in Boolean attribute is Boolean type. However, when the feature is $x_i = 0$, the feature term is not added to the calculation of the conditional probability. At this time,

$$P(\overline{X} \mid C) = \prod_{i=0}^{m-1} p(t_i \mid C)^{x_i}. \tag{6}$$

And the Laplacian smoothing method is different.

2.3. Establishment of Word Sense Disambiguation Model. Most machine learning methods have been applied to the field of word sense disambiguation, which can be divided into discriminative models and generative models according to different model learning methods. In terms of accuracy and efficiency of disambiguation, the Bayesian model has a good balance, and the good robustness of the Bayesian model is also a key point for many word sense disambiguation models to use the model. The English combination ambiguity resolution framework is shown in Figure 2.

According to the characteristics of Chinese word sense disambiguation, this paper uses the sliding word window method to open the word window according to the position of the target ambiguous word, uses the Dice coefficient method to obtain the semantic information of the feature words in the word window as the disambiguation feature, and constructs a feature vector set. The feature vector set is applied to the Bayesian model to obtain a semantic sense disambiguation model based on semantic knowledge.

In the language environment where the ambiguous word is located, the final interpretation of the ambiguous word is determined based on the maximization of the posterior probability, if

$$P(S_j \mid \text{Context}) \geq P(S_i \mid \text{Context}), \quad (i = 1, 2, \ldots, m). \tag{7}$$

The semantics of the ambiguous word w is S_j. Among them, Context is the context in which the ambiguous word w is located. Generally, Context is composed of word units on both sides of w, which provide necessary guidance information for the disambiguation process. The ambiguous word w has m semantic categories $S_1, S_2, \ldots, S_m$. In Context, its true semantic category is S_j. $P(X)$ is the probability of X appearing. In the context, if the probability of the ambiguous

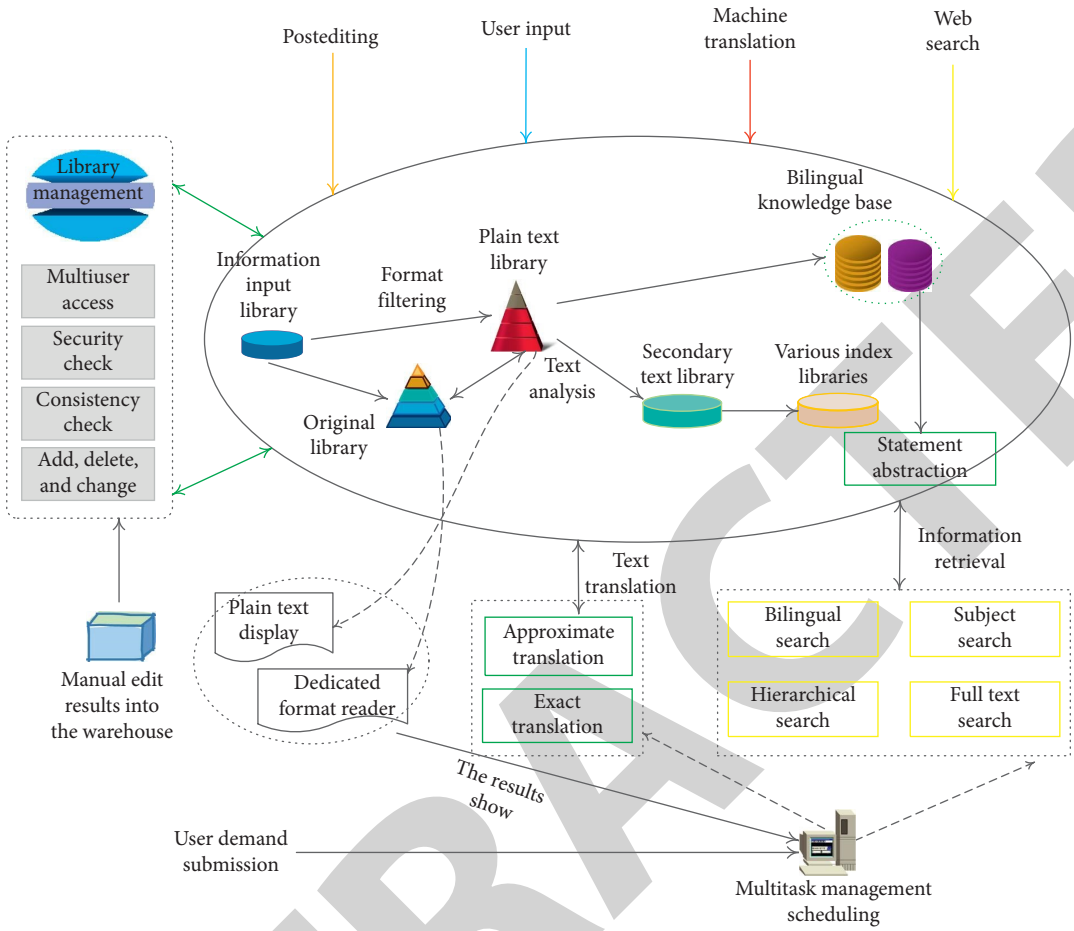

Figure 1: Network bilingual information processing system for machine translation.

word w taking the semantic category S_j is greater than any other semantic category S_i ($j \neq i$), the semantic category of the word w should be determined as S_j. Bayesian decisions have the smallest probability of error. For the ambiguous word w, it has m semantics $S_1, S_2, \ldots, S_m$, and the context disambiguation features are FL and FR, respectively. Among them, FL and FR are semantic class codes. The semantic code of "Synonym Cilin" is divided into three layers: FL = fl1fl2fl3; FR = fr1fr2fr3.

The process of word sense disambiguation based on semantic knowledge is as follows:

(1) We analyze the corpus to obtain sentence information, segmentation information, and part-of-speech information.

(2) We use the sliding word window method to find the position of the target ambiguous word and use it as the center to open the word window on both sides to obtain the word window segmentation.

(3) We use "Synonyms Word Forest" as a semantic dictionary to obtain the semantic code set of word segmentation in the word window.

(4) We use the Dice coefficient method to determine the word segmentation code. The feature vector set is constructed by three different feature extraction methods: first-level semantic code, three-level semantic code, and morphological information.

(5) We construct a word sense disambiguation model based on semantic knowledge by using the feature vector set extracted from the training corpus.

(6) We disambiguate the test corpus by constructing a good word sense disambiguation model. Three different tests are performed on different feature vector sets.

2.4. Construction and Analysis of Numerical Simulation Experiments. We select 10 more polysemous words for comparison. These words include dichotomous vocabulary and multiple vocabularies. We made statistics on their various data, and the data situation is shown in Figure 3.

We obtain the experimental corpus and preprocess the corpus. The preprocessing part mainly includes analyzing the corpus and locating the target ambiguous words. The sliding word window method is used to obtain the lexical information set near the target ambiguous words. The feature processing part mainly includes acquiring the semantic knowledge set and determining the semantic knowledge by the Dice coefficient method. Semantic codes can be divided into three categories: large, medium, and small.

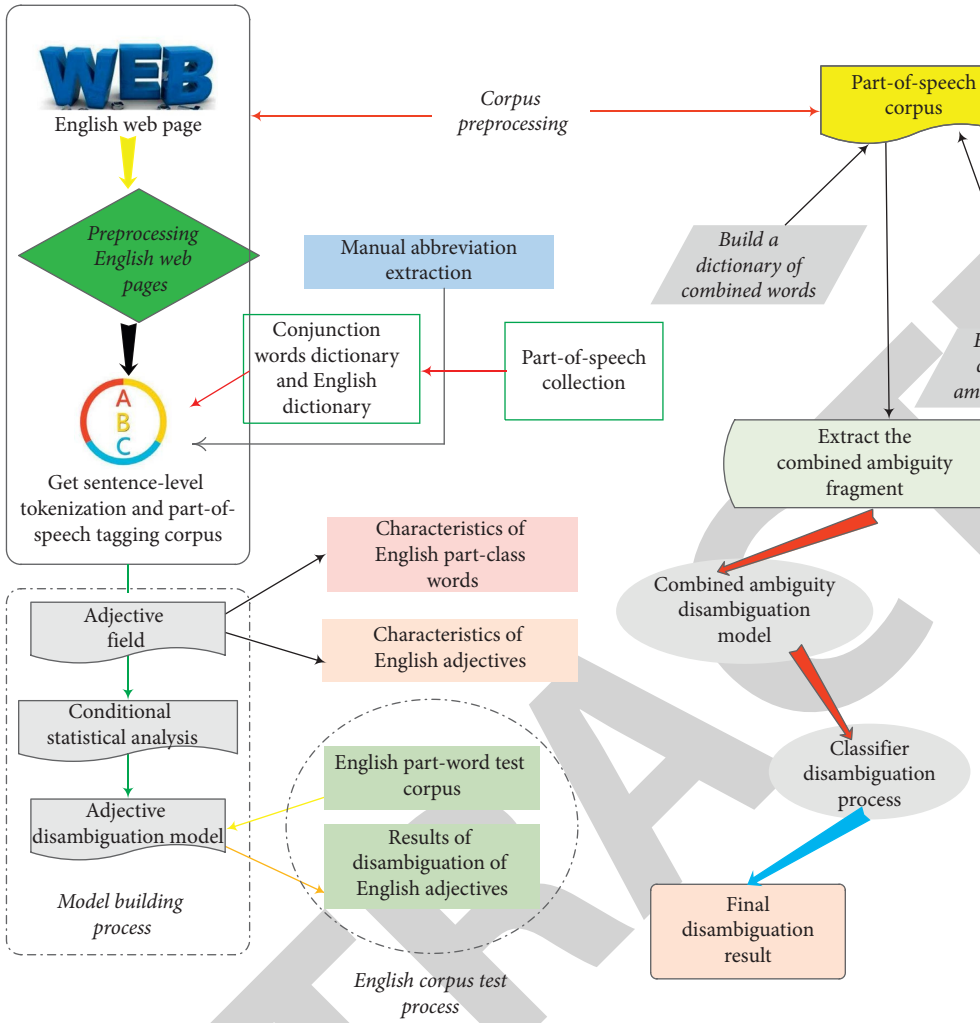

Figure 2: English combination ambiguity resolution framework.

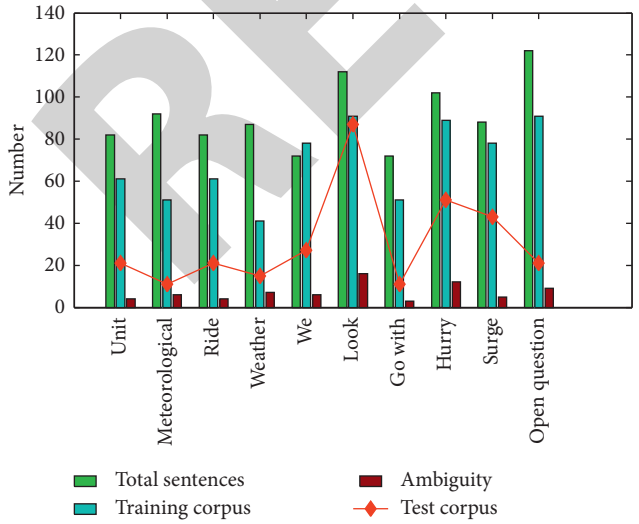

Figure 3: Detailed corpus statistics.

The word meaning encoding extraction method obtains four semantic class codes. The first-level encoding extraction method obtains three first-level semantic codes, which are "D, H, and J." Although "Synonym Cilin" has a more precise five-layer structure, which is the "Synonym Cilin" expansion board, this version of the semantic dictionary is different from the purpose of "Synonym Cilin" at first, but it is closer to the definition dictionary. If the extended version is used for implementation, the three-layer structure code of the word is almost the same as the word form information of the word, which results in no essential difference between the two feature vector sets.

From the above coding set, the most probable semantic knowledge information in the target word window is calculated and obtained according to the Dice coefficient method. Using three different feature extraction methods, three different semantic knowledge information feature vector sets are formed, and a semantic-based disambiguation model is constructed. Finally, the test corpus is used to perform experiments in the disambiguation model, and the accuracy of the models formed by three different feature extraction methods is recorded for the final comparison.

In the experimental stage, in order to measure the impact of training corpus and test corpus distribution on word sense disambiguation, relevant statistics were performed on the training data set. In a given corpus, the number of occurrences of each corpus is counted, and the results are shown in Figure 4.

It can be seen from Figure 4 that the interpretation of each vocabulary of the training corpus is about three times that of the test corpus. This corpus division is also in line with the simple cross-validation in statistical learning. In order to compare the results of the word sense disambiguation method proposed in this paper, we test the efficiency of the method through comparative experiments. Figure 5 lists the disambiguation accuracy of the two different methods.

3. Unsupervised Professional Field Segmentation Ambiguity Resolution Method

3.1. Analysis of Word Segmentation Ambiguity in Professional Field. The words that cause the ambiguity of the professional domain corpus and the general domain corpus are mainly short words. In the Fish corpus, two-word words accounted for 79%, and three-word words accounted for 13%; in the Bird corpus, two-word words accounted for 81%, and three-word words stood for 14%. Compared with the general field segmentation ambiguity, the composition of professional field segmentation ambiguity is slightly more complicated. In the professional domain corpus, not only ordinary words and ordinary words but also ordinary words and domain words cause segmentation ambiguity. For example, the term "neutralization" (neutralization reaction, the exchange of acids and bases with each other, and the reaction of salt and water) is often used in chemistry. In common sentences, "remaining in the measuring cup and the test strip" is often used.

The professional field lacks labeled training corpus and professional field knowledge, cannot count *N*-gram grammatical information, and cannot quickly generate professional field word segmentation ambiguity resolution rules; the professional field words and common words in different sentences cannot be generalized. Based on the above reasons, the general Chinese word segmentation ambiguity resolution method is not suitable for solving the problem of word segmentation ambiguity in professional fields.

Chinese word segmentation mainly includes two types of word segmentation ambiguity: cover ambiguity and overlap ambiguity. Statistical analysis of the number and phenomenon of the two types of ambiguities in the corpus shows that the number of coverage ambiguities is small and the number of overlap ambiguities is large.

Based on the vocabulary, the method of "FMM + back word" can be used to find the overlapped ambiguities that may exist in the test corpus. A set of overlapping strings owned by an overlapping type cut is called an overlapping string chain, and its number is called a chain length. For example, "Secondary" consists of two words overlapping "Second" and "Secondary." The set of overlapping strings is {"time"}, and the chain length of the overlapping field is 1. In

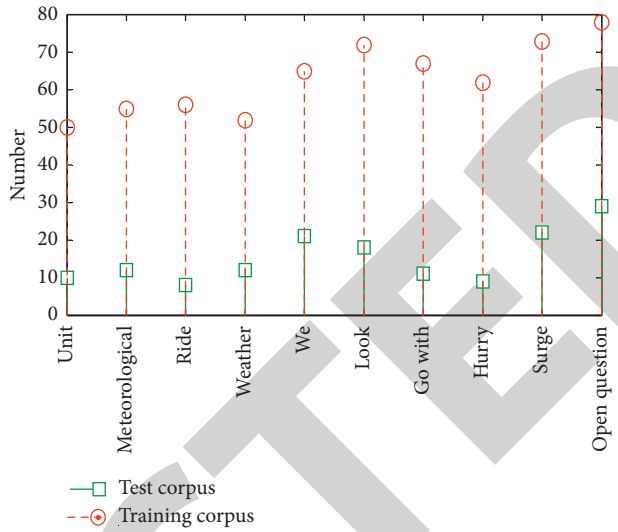

Figure 4: Corpus distribution.

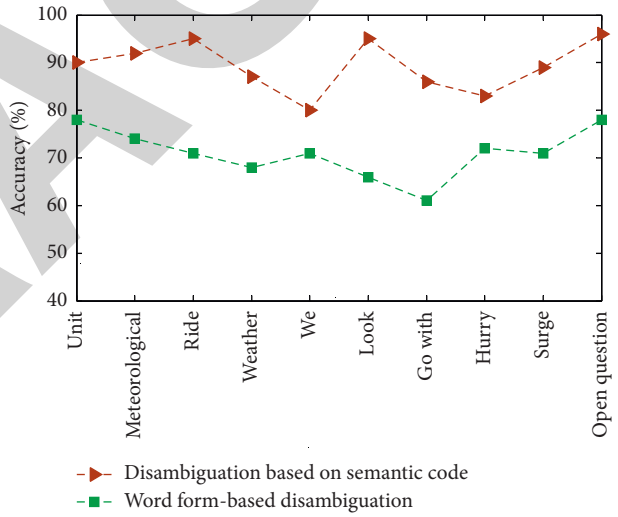

Figure 5: Comparison of word sense disambiguation results.

Table 1: MOAS chain length statistics.

	Fish corpus					Bird corpus				
Chain length	1	2	3	4	5	1	2	3	4	5
Number	741	269	8	3	2	1189	422	40	4	2
Share	0.73	0.26	0.01	0.02	0.01	0.72	0.25	0.02	0.02	0.01

the "molecule," "combination," "synthesis," "component," and "molecule" all form words, and the chain length of the overlapping field is 3.

Using the word fallback method, based on the Fish vocabulary and Bird vocabulary, we find possible overlap ambiguities in all test corpora and obtain MOAS (Maximal Overlapping Ambiguity String). We classify MOAS in Fish and Bird corpus according to chain length. The statistical results of MOAS chain length are shown in Table 1.

3.2. Unsupervised Professional Field Segmentation Ambiguity Resolution Method. Some methods in unsupervised word segmentation are often used to determine the likelihood of a string forming a word in the current corpus. Mutual information can be used to quantitatively estimate the binding force between two Chinese characters. The larger the mutual information, the higher the closeness of the combination of the two Chinese characters; the smaller the mutual information, the lower the tightness of the combination. The calculation formula is

$$\mathrm{mi}\,(x, y) = \log_2 p(x, y) - \log_2 p(x) - \log_2 p(y). \quad (8)$$

Mutual information can only express the combining power of two Chinese characters. When it encounters a word composed of three or more multicharacters, it cannot express it. Based on mutual information, this paper changes the multiword mutual information. When a word is composed of multiple words, the mutual information of two adjacent words is calculated first, and then the average value is taken. The extended mutual information is

$$\mathrm{mi}'\left(x_{1,\ldots,n}\right) = (n-1)^{-1} \sum_{i=1}^{n} \mathrm{mi}\,\left(x_i, x_{i+1}\right). \quad (9)$$

Boundary entropy is also one of the important methods to measure whether a string is a word often used in unsupervised word segmentation. The formula is

$$h\left(x_{i,\ldots,j}\right) = -\sum_{x \in V} \log\, p\left(x \,\middle|\, x_{i,\ldots,j}\right) p\left(x \,\middle|\, x_{i,\ldots,j}\right), \quad (10)$$

when x is x_{i-1}, x is the set of left-adjacent words, and $h(x_{i,\ldots,j})$ is called the left-handed conditional entropy and when x is x_{j+1}, x is the set of right-adjacent words, and $h(x_{i,\ldots,j})$ is the right-handed conditional entropy. The greater the left and right entropy of a string, the more likely it is a word.

String boundary entropy can disambiguate segmentation in two ways. The first is to judge the possibility of string formation. The second is to judge the degree of separation between words. When judging the possibility of forming a string of words, the left entropy and right entropy of each string in the test corpus are statistically calculated, and the average value of the left entropy and right entropy of each word is taken as a measure. When the string *abc* has two splitting methods, *ab*/*c* and *a*/*bc*, the string boundary entropy used by the strings *ab* and *bc* is

$$\begin{aligned} h_{\mathrm{avg}}(ab) &= 0.5\,[hr(ab) + hl(ab)], \\ h_{\mathrm{avg}}(bc) &= 0.5\,[hr(bc) + hl(bc)]. \end{aligned} \quad (11)$$

To determine the degree of separation between strings, we use the following formula:

$$h_{\mathrm{separate}}\left(x_{i,\ldots,j}\, y_{i,\ldots,j}\right) = 0.5\left[hr\left(x_{i,\ldots,j}\right) + hl\left(y_{i,\ldots,j}\right)\right]. \quad (12)$$

When the string *abc* has two splitting methods, *ab*/*c* and *a*/*bc*, the degree of separation between *ab* and *c* can be expressed as

$$h_{\mathrm{separate}}\left(\frac{ab}{c}\right) = hl(c) + hr(ab). \quad (13)$$

The degree of separation between *a* and *bc* can be expressed as

$$h_{\mathrm{separate}}\left(\frac{a}{bc}\right) = hl(bc) + hr(a). \quad (14)$$

Boundary entropy can be used not only to judge the possibility of forming a string of multiple words but also to judge the possibility of a single word becoming a "single word."

In the general field segmentation, the more frequently a word in the training corpus becomes a single word, the more likely it is that the word will become a single word in the test corpus. In the string *abc* with a length of 3, whether it is split into *ab*/*c* or *a*/*bc*, you can determine who has *a* and a word more frequently in the training corpus; if *a* is more frequent than a single word *c*, the segmentation result is *a*/*bc*. There is no training corpus for word segmentation in professional fields, and it is impossible to count the probability of single words. When encountering a string *abc* with a length of 3, you can use havg to determine which word in the test corpus is more likely to become a single word.

3.3. Numerical Simulation Experiment

3.3.1. Experimental Corpus and Settings. The test corpus in the experiment is the corpus of fish and birds in the Biovolume of China Encyclopedia. The participle answer of the corpus is marked with Peking University participle standard.

The choice of experimental corpus follows two basic principles: First, it only observes the ambiguity of the segmentation in the domain and eliminates the interference of the OOV (Out of Vocabulary) problem on the effect of the segmentation. All words needed for the test corpus are added to the vocabulary. The second is the maximum possible simulation of actual application. For this experiment, large-scale vocabularies, common domain vocabularies, and network vocabularies are added to the experimental vocabulary.

Taking string frequency judgment as an example, the experiment considers two cases:

(1) MOAS consists of three characters *abc*, which can be divided into *ab*/*c* and *a*/*bc*. If Freq (*ab*) > Freq (*bc*), it is divided into *ab*/*c*.

(2) MOAS is more than 3 characters, such as *abcd*; if there are two forms of *abc*/*d* and *ab*/*cd*, when Freq (*abc*) > Freq (*cd*), it is divided into *abc*/*d*; when Freq (*cd*) > Freq (*abc*), you need to consider again. After dividing *cd* into words, *ab* is still a word in the dictionary. If *ab* does not form a word, *abc*/*d* is still used; if *ab* is formed, it is *ab*/*cd*.

When mutual information is used as the evaluation criterion, the experimental procedure is the same as the frequency of the character string, and Freq is changed to mi. When taking the possibility of forming a string boundary entropy as the evaluation criterion, the experimental steps are the same as the string frequency experiments, and Freq is changed to havg.

Table 2: Treatment of 10 experiments.

Experiment number	Processing situation
1	Baseline segmentation results without any ambiguity
2	Baseline segmentation results without any ambiguity
3	Freqword first uses the vocabulary to segment the test corpus, then counts the word frequency information in the test corpus after segmentation, and finally uses the new statistical word frequency information to segment the test corpus
4	Freq only uses string frequency to process ambiguous segmentation results
5	Freq + zi_havg uses a single-word boundary entropy judgment when MOAS consists of 3 characters, and more than 3 characters use Freq judgment
6	mi uses only mutual information
7	mi_zi + havg uses single-word boundary entropy when MOAS is composed of 3 characters, and mutual information is used for more than 3 characters
8	word_havg segmentation results use string boundary entropy to determine the likelihood of word formation
9	word_hseparate uses only string boundary entropy to judge string separation results
10	word_havg + zi_havg uses the boundary entropy of a single word when MOAS is composed of 3 characters and uses the boundary entropy of a string to determine more than 3 characters

When the separation degree of the string boundary entropy is used as the evaluation standard, the experimental steps are the same as the string frequency experiments, but the evaluation rule is changed to determine hseparate (ab/c).

3.3.2. Experimental Results and Analysis. In order to analyze only the effect of word segmentation ambiguity, based on the Fish vocabulary and Bird vocabulary without OOV, the FMM word segmentation method was used to test the corpus segmentation of Fish and Bird, respectively. The segmentation result without any ambiguity is used as the baseline of the experiment. The processing conditions corresponding to the 10 experiments are shown in Table 2.

The data in Figures 6 and 7 show that, after only using string frequency to resolve the ambiguity problem, the F1 value of the word segmentation result has been significantly improved, the Fish result has increased by 0.48 percentage points, and the Bird result has increased by nearly 0.9 percentage points. Mutual information and string boundary entropy behave differently on different corpora. Mutual information has the best effect on the Fish corpus, with an increase of nearly 0.8 percentage points; on Bird corpus, the best effect is obtained by mixing the string boundary entropy with the single-word boundary entropy, increasing by 1.2 percentage points. The Fish-based corpus of m_i-based segmentation surpasses Topline.

The experimental results of word_hseparate show that using this method to resolve the ambiguity of the word segmentation, the effect of the word segmentation decreases. The reason is that this method is equivalent to a local optimal solution, which results in many erroneous results.

In Bird's corpus, mixed word boundary entropy is better than single-word entropy. However, the results of the Fish corpus are different. The statistical data of single-word boundary entropy comes from the test corpus. If some words in the test corpus often appear at various positions of the word, the single-word boundary entropy of the word will also be high. For example, the "fish" in the Fish corpus appears very frequently, and the single-word boundary entropy of "fish" is very high, but the possibility of becoming a single word is very small. It is generally used as the beginning and end of a word, such as "fish bait" and "fish." There are many such words in the Fish corpus, and the

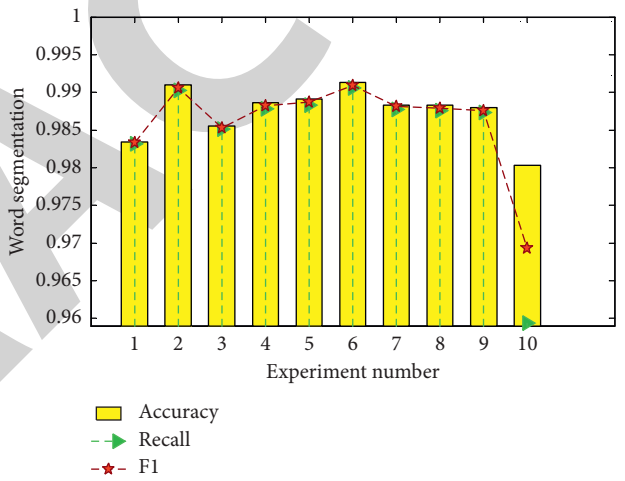

Figure 6: Various word segmentation results of Fish without OOV.

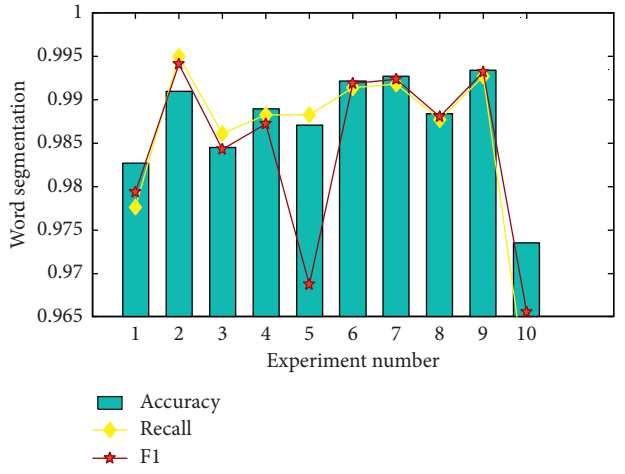

Figure 7: Bird's various word segmentation results without OOV.

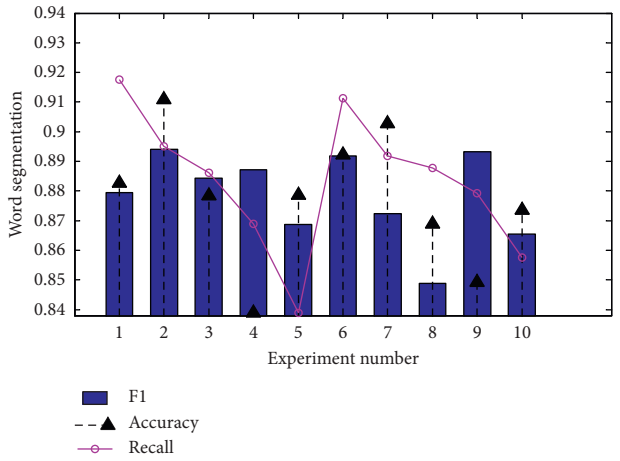

Figure 8: Various word segmentation results of Fish with OOV.

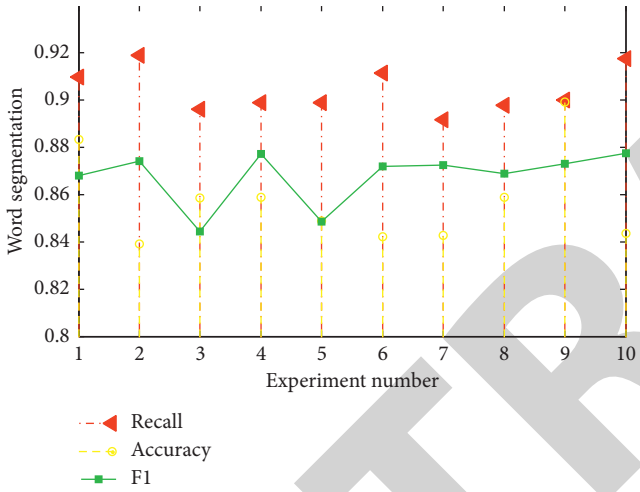

Figure 9: Various word segmentation results of Bird with OOV.

single-word boundary entropy does not improve the Fish corpus. There are fewer such words in birds, and the accuracy of the single-word boundary entropy is higher, which has an effect on the result of word segmentation.

The boundary entropy of the word in the test corpus is subtracted from the boundary entropy of the word in the vocabulary, and the difference obtained by subtraction is the final single-element boundary entropy. The boundary entropy of some words after subtraction is negative. In this experiment, the single-word boundary entropy is only used for comparison, it is not used for other operations, and no other processing is performed on negative boundary entropy.

To further verify the validity of the results, the word segmentation vocabulary was changed. The professional words required by the test corpus are halved randomly, so that 879 OOVs (word types) in Fish cannot be recognized, and 1468 OOVs (word types) in Bird cannot be recognized.

The data in Figures 8 and 9 show that, in the presence of OOV, using string frequency, mutual information, and boundary entropy to process ambiguity, the segmentation results are better than baseline. Fish's best results and Bird's best results are improved. Among all ambiguity processing methods, the performance of mutual information is the most stable.

4. Conclusion

According to the constructed feature vector set and through comparison, this paper chooses a naive Bayesian model with better efficiency and accuracy. A word sense disambiguation classifier is established based on the characteristics of semantic codes. A comparison experiment is performed through three different feature vector sets to analyze the performance of the word sense disambiguation classifier based on semantic information. The experimental results show that the accuracy of disambiguation of Bayesian word sense disambiguation model based on semantics is greatly improved compared with the traditional word sense disambiguation model. The general domain segmentation ambiguity resolution method requires professional knowledge and training corpus and is not suitable for solving the professional domain segmentation ambiguity problem. In this paper, the frequency, mutual information, string boundary entropy, and single-word boundary entropy of the test corpus are used as evaluation criteria to solve the problem of word segmentation ambiguity in professional fields. Experiments show that the three evaluation standards can solve the problem of word segmentation ambiguity in professional fields to varying degrees. Among them, the segmentation words using mutual information have the best results and perform stably. The word segmentation method combined with unsupervised learning in the professional field is simple and easy, and it can effectively reduce the problem of word segmentation ambiguity in the test corpus and improve the word segmentation effect in the professional field. However, natural language processing is increasingly a large-scale corpus. Therefore, how to deal with various types of ambiguity fields and improve the speed of word segmentation needs further study.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] L. Deng and D. Li, "Multimedia data stream information mining algorithm based on jointed neural network and soft clustering," *Multimedia Tools and Applications*, vol. 78, no. 4, pp. 4021–4044, 2019.

[2] W. S. Grant and L. Itti, "Learning invariant features in modulatory networks through conflict and ambiguity," *Neural Computation*, vol. 31, no. 2, pp. 344–387, 2019.

[3] Y. Graham, T. Baldwin, A. Moffat, and J. Zobel, "Can machine translation systems be evaluated by the crowd alone," *Natural Language Engineering*, vol. 23, no. 1, pp. 3–30, 2017.

[4] P. Iswarya and V. Radha, "Adapting hybrid machine translation techniques for cross-language text retrieval system,"

Journal of Engineering Science and Technology, vol. 12, no. 3, pp. 648–666, 2017.

[5] M. M. A. Shquier and K. M. Alhawiti, "Fully automated Arabic to English machine translation system: transfer-based approach of AE-TBMT," *International Journal of Information and Communication Technology*, vol. 10, no. 4, pp. 376–391, 2017.

[6] M. Howlett, "Moving policy implementation theory forward: a multiple streams/critical juncture approach," *Public Policy and Administration*, vol. 34, no. 4, pp. 405–430, 2019.

[7] S. Zannettou, M. Sirivianos, J. Blackburn, and N. Kourtellis, "The web of false information," *Journal of Data and Information Quality*, vol. 11, no. 3, pp. 1–37, 2019.

[8] E. M. Garcia, C. Creus, C. España-Bonet, and L. Màrquez, "Using word embeddings to enforce document-level lexical consistency in machine translation," *The Prague Bulletin of Mathematical Linguistics*, vol. 108, no. 1, pp. 85–96, 2017.

[9] M. Bouchakwa, Y. Ayadi, and I. Amous, "Multi-level diversification approach of semantic-based image retrieval results," *Progress in Artificial Intelligence*, vol. 9, no. 1, pp. 1–30, 2020.

[10] J. Joseph and V. Gaba, "Organizational structure, information processing, and decision-making: a retrospective and road map for research," *Academy of Management Annals*, vol. 14, no. 1, pp. 267–302, 2020.

[11] S. Mall and U. C. Jaiswal, "Survey: machine translation for Indian language," *International Journal of Applied Engineering Research*, vol. 13, no. 1, pp. 202–209, 2018.

[12] A. Rezapour, S. M. Fakhrahmad, and M. H. Sadreddini, "Applying various distance functions and feature extraction schemes to ambiguity resolution," *Intelligent Data Analysis*, vol. 22, no. 3, pp. 617–638, 2018.

[13] Y. F. Hassan, "Rough set machine translation using deep structure and transfer learning," *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 6, pp. 4149–4159, 2018.

[14] P. Xiu and L. Xeauyin, "Human translation vs machine translation: the practitioner phenomenology," *Linguistics and Culture Review*, vol. 2, no. 1, pp. 13–23, 2018.

[15] G. Kotzé, V. Vandeghinste, S. Martens, and J. Tiedemann, "Large aligned treebanks for syntax-based machine translation," *Language Resources and Evaluation*, vol. 51, no. 2, pp. 249–282, 2017.

[16] F. Maniez, "An appraisal of recent breakthroughs in machine translation: the case of past participle-based compound adjectives in ESP," *ASp*, vol. 2017, no. 72, pp. 29–48, 2017.

[17] A. Benabdallah, M. A. Abderrahim, and M. E.-A. Abderrahim, "Extraction of terms and semantic relationships from Arabic texts for automatic construction of an ontology," *International Journal of Speech Technology*, vol. 20, no. 2, pp. 289–296, 2017.

[18] J. Ive, A. Max, and F. Yvon, "Reassessing the proper place of man and machine in translation: a pre-translation scenario," *Machine Translation*, vol. 32, no. 4, pp. 279–308, 2018.

[19] M. Kathuria, C. K. Nagpal, and N. Duhan, "A fuzzy logic based synonym resolution approach for automated information retrieval," *International Journal on Semantic Web and Information Systems*, vol. 14, no. 4, pp. 92–109, 2018.

[20] K. Cui and X. Jing, "Research on prediction model of geotechnical parameters based on BP neural network," *Neural Computing and Applications*, vol. 31, no. 12, pp. 8205–8215, 2019.

[21] M. Chen, S. Lu, and Q. Liu, "Uniform regularity for a Keller-Segel-Navier-stokes system," *Applied Mathematics Letters*, vol. 107, Article ID 106476, 2020.

[22] J. Kapočiūtė-Dzikienė, A. Davidsonas, and A. Vidugirienė, "Character-based machine learning vs. language modeling for diacritics restoration," *Information Technology and Control*, vol. 46, no. 4, pp. 508–520, 2017.