Hindawi
Advances in Multimedia
Volume 2022, Article ID 4421976, 10 pages
https://doi.org/10.1155/2022/4421976

Research Article

English Syntactic Analysis and Word Sense Disambiguation Strategy of Neutral Set from the Perspective of Natural Language Processing

Chaohui Liang and Jiling Shang

Department of Basic Teaching, Zhengzhou Railway Vocational and Technical College, Zhengzhou 450018, China

Correspondence should be addressed to Chaohui Liang; 10161@zzrvtc.edu.cn

Received 9 May 2022; Revised 27 May 2022; Accepted 14 June 2022; Published 8 August 2022

Academic Editor: Qiangyi Li

In order to improve the effect of English semantic analysis, under the support of natural language processing, this paper analyzes English syntactic analysis and the word sense strategy of the neutral set and solves the parameters through data training, so as to solve the probability distribution of the maximum entropy model of each order. Moreover, by comparing the prediction probability of the model to the judgment mode with the experimental data, it is found that the first-order maximum entropy model (independent model) is quite different from the data. Therefore, when judging data in English semantics, we cannot only consider the influence of second-order correlations but should also consider higher-order correlations. The research results of the simulation experiment show that the English syntactic analysis and the word sense disambiguation strategy of the neutral set proposed in this paper from the perspective of natural language processing are very effective.

1. Introduction

One of the trends in linguistics research is the increasing emphasis on using data to illustrate problems. Thanks to the development of acoustic technology, phonetics can collect data more precisely. Moreover, combined with statistical analysis methods, phonetics research has matured quantitative analysis capabilities. However, the traditional quantitative research on grammar mainly counts the frequency of occurrence of a certain type of grammar unit or a specific format, and its explanatory power is limited. Therefore, grammar research must take new approaches. Natural language processing is one of the important directions of applied linguistics.

Through parsing sentence syntactic analysis, the internal structure information between words can be generated, which provides convenience for information acquisition, question answering system, and machine translation [1]. Syntactic analysis can be divided into dependency structure syntactic analysis and phrase structure syntactic analysis. The current mainstream method is a syntactic analysis method with supervised learning by using a large-scale artificially annotated corpus [2]. Due to the lack of morphological changes in speech and a large number of concurrent words, the accuracy of its analysis does not reach the accuracy of English syntactic analysis, and it is still a research hotspot [3]. The performance of supervised syntactic analysis algorithms depends to a large extent on manually labeled data. The quality and scale of the labeled data will seriously affect the performance of the algorithm. Manually labeling data is a very heavy task and requires professional linguistic knowledge. Therefore, it is not easy to obtain high-quality data [4]. Since the labeling rules of each labeling corpus are different and cannot be directly combined, how to effectively use the existing labeling data to expand the training data is also a research hotspot [5].

With the in-depth development of neural networks and deep learning in the field of natural language processing, the method of dependency parsing based on the neural network has developed rapidly. The idea of the transition-based

dependency parsing method is to reduce the complex task to the simple task of predicting the next parsing action and then use the classifier to perform a greedy search of the optimal sequence [6]. Reference [7] applies a simple neural network to dependency parsing, uses a neural network to score decision actions, and uses a new cubic activation function to model the interaction between inputs. Reference [8] proposed to use of a sequence-to-sequence neural network, the stack LSTM, to implement the state-of-the-art transition-based dependency parser at the time. Reference [9] conducts global normalization training for transfer-based neural network models to overcome the label bias problem caused by local normalization models. Based on the transition-based dependency parsing system, literature [10] proposed a method for Telugu dependency parsing by using the minimum feature function represented by context vectors to replace the language feature templates used in the past. The graph-based dependency parsing method treats the dependency parsing problem as the problem of finding the maximum spanning tree from a fully directed graph. The score of a dependency tree is obtained by accumulating the scores of several subtrees that constitute the dependency tree, and the exact search for the maximum spanning tree is realized on the global optimization model. Reference [11] proposes a feature extraction method for the BiLSTM encoder jointly trained with the parser and applies it to the transfer-based parser and graph-based parser. The method works in both dependency parsers. There is a good analytical effect. Reference [12] proposed to use a more regularized parsing method than the BiLSTM encoder based on the graph-based dependency parsing method, that is, using a double affine classifier to score dependency arcs and dependency labels. Reference [13] proposed a dependency parsing method, developed five Thai dependency parsing algorithms from transition-based dependency parsing, and developed two Thai dependency parsing algorithms from graph-based dependency parsing. The algorithms used pretrained models to learn character embedding and used BiLSTM to process bidirectional features to effectively overcome word order problems. Reference [14] performed well in both transfer-based and graph-based methods. Reference [15] studied the effect of encoding based on a bidirectional long short-term memory network (BiLSTM) on the scoring step. Reference [16] studied the influence of different structural constraints in the decoding stage on the results of dependency analysis, and the results show that by global input modeling, even ignoring some output structures can get good results. Although transfer-based methods can use rich features, they all use local search strategies, which are prone to error propagation. Graph-based methods have a smaller feature range but can perform a global search, can handle long-range dependencies and nonprojective phenomena, and are more accurate than transfer-based methods. Dependency syntax analysis has high requirements on sentence structure, and the distribution of real scenes and labeled data is obviously different. If manual labeling is used, it will be difficult to label and easy to generate wrong data. Reference [17] used a thesaurus-based substitution method to augment data in the data augmentation stage. Reference [18] uses back translation and random noise injection to augment the unlabeled text. Reference [19] augments the data by using text surface transformation, such as the change between abbreviations, from "Itisawesome" to "It'sawesome." A simple and effective image enhancement technique is proposed in [20]. The idea is to combine two random images in a certain proportion in a mini-batch to generate synthetic instances for training. This idea was subsequently applied in the field of NLP.

Aspect-level sentiment analysis includes two subtasks, aspect extraction and aspect-level sentiment classification, where the effectiveness of aspect extraction is the premise to ensure the accuracy of fine-grained sentiment analysis. Aspects can be divided into explicit and implicit aspects. The explicit aspect refers to the evaluation aspect of words directly contained in the sentence. Taking the evaluation in the restaurant field as an example, most techniques in current aspect-level sentiment analysis research are about explicit aspects,\ and implicit aspects are rarely mentioned, that is, most of the models focus on extracting the explicit aspects in the first sentence above, while ignoring the extraction of the implicit aspects in the second sentence, losing some useful information in the data, and not guaranteeing the integrity of the data analysis. Implicit aspects account for about 30% of sentences. Implicit aspect extraction technology not only plays an important role in improving the accuracy of aspect extraction but also plays an important role in the comprehensiveness and integrity of fine-grained sentiment analysis research. At present, implicit aspect extraction techniques are mainly divided into three categories as follows: relational inference, topic clustering, and taxonomy. The relational inference method is based on the corresponding relationship between explicit aspect words and opinion words, and mining implicit aspects according to opinion words. It mainly introduces different methods to improve the rules on the basis of the cooccurrence frequency analysis method or association rule method.

Syntactic analysis has a wide range of applications in machine translation, question answering systems, information extraction, and speech synthesis systems and has always been one of the keys and difficult points of natural language processing research. Automatic syntactic analysis refers to analyzing the relationship between the grammatical units contained in a sentence and these grammatical units according to a certain grammatical system under the condition of a given word sequence and converting the linear word sequence into a grammatical tree with a hierarchical structure. Syntactic analysis and semantic analysis are two important aspects of natural language processing, and there is a close relationship between them. Therefore, to understand a sentence, it is necessary to not only analyze its syntactic structure but also understand the semantics of each word in it. Research has shown that tasks at two levels, syntactic analysis and semantic analysis, can help each other. According to the characteristics of Chinese, the combination of syntactic analysis and semantic analysis is helpful to the processing of Chinese. On the other hand, the difficulty of natural language processing lies in the problem of ambiguity. In syntactic analysis, ambiguity in a syntactic structure needs

to be resolved, and word sense disambiguation in semantic analysis needs to deal with the ambiguity of polysemous words according to the context. The solution of either of these two ambiguity problems will greatly help the other task, and if they can be solved in the same process, the connection between them can be fully exploited to benefit both.

With the support of natural language processing ideas, this paper analyzes the English syntactic analysis and the word sense of the neutral set and promotes the effect of English semantic analysis.

2. Syntactic Analysis Based on Natural Language Processing

2.1. Probabilistic Model for Orientation Discrimination Based on the Maximum Entropy Method. In the real natural environment, organisms always receive many external stimuli at the same time. Thus, how humans respond to multiple stimuli has long been a central question in linguistics. Among them, an important aspect is the effect of the number of stimuli on the response.

In human thinking, judging whether things are the same is the most basic performance. For example, when we see two lines, we will judge whether the angle and length of the two lines are the same according to the characteristics of the lines (such as angle and length). Similarity judgments are used in many psychophysical tasks, such as singularity finding and causal inference.

A natural idea is how people's judgment of the similarity of things will change when faced with multiple environmental attribute stimuli (such as the number, length, and color of lines).

We consider an experiment for similarity judgment as follows: judging whether the line directions are the same. In order to simplify the experiment, four-line attributes are considered as follows: number, length, color, and thickness, denoted as $X = (x_1, x_2, x_3, x_4)$, X is a display mode (environmental mode), and the relationship between human judgment and different display modes is studied. Suppose there are two values for each attribute, so there are 2^4 different patterns, and the human judgment is the same or not, so there are 32 different combinations, which is more complicated than before. At the same time, in order to meet the more general situation and not make assumptions such as independence, it is very important to select a suitable model, and the maximum entropy model can better solve this problem.

The maximum entropy model was first used in word classification, and later, due to its excellent performance, it has been widely used in biology, computer science, probability theory, statistics, cognitive linguistics, and other disciplines. Schneidman et al. first applied the principle of maximum entropy in the coding of neuron clusters, and the constructed model could contain the correlations between neurons. Globerson et al. used the principle of minimal informatization to analyze the contribution of the correlation model to the encoded information. On this basis, Marre et al. studied the application of the maximum entropy model in the coding of neuron clusters, focusing on the influence of pairwise spatial correlations on neuron clusters. Cayco Gajic et al. studied the correlation between the three neurons (triplet correlations).

The maximum entropy principle is applied to the language judgment problem of similar items, and the distribution of human language judgment under different display modes is obtained. Moreover, this paper finds out the implicit correlation and uses the change of entropy to intuitively quantify the performance of the model to the test data.

2.2. Higher Order Maximum Entropy Model and Direction Discrimination. Correlative experiments are designed based on psychophysical experiments that judge whether a group of items is the same. In the judgment test, there are a total of n subjects, and the task is to judge whether the directions of the lines appearing in the display are all the same. In the experiment, each subject did 1000 repetitions of the experiment. The experimental process is shown in Figure 1. Among them, the environmental conditions that affect the subjects' judgment are as follows: the number of lines (2 or 4), the length (1 or 2), the width (0.2 or 0.5), and the color (red or gray). In each trial, the probability of the lines being all the same and different is the same. With the same orientation, the orientation of the lines is drawn from the uniform distribution U (1,180). In the case that the directions are not all the same, the angle θ_i of the i-th line is randomly selected from the uniform distribution U (1,180), and the other lines have the same direction and are drawn from the normal distribution $N(\theta_i, \sigma^2)$, where $\sigma = 10^\circ$ is taken. The attribute conditions of the lines are relatively weakly correlated and appear randomly.

In each trial, the first is the preparation phase, a cross will appear in the display, and when the subject is ready, press the space bar to enter the trial phase. At this point, lines with different ambient patterns (number, color, thickness, length) appear, and each line has the same pattern. Subjects need to make a judgment as follows: when they are in the same direction, then enter the judgment stage. If it is the same, the subject presses "1," and if it is different, the subject presses "0."

In general, the maximum entropy model is quite different from the traditional model. Traditional methods need to make some simple assumptions first and then perform calculations. Naturally, this inevitably produces some errors in the details. However, the maximum entropy method can only rely on the existing performance of the experimental data without making additional assumptions, so as to retain as much information as possible in the data.

We set $x_i = 1$ or 0, indicating that the i-th condition occurs or does not occur. The vector $\vec{x}$ represents a judgment mode. Thus, if we have N conditions, there will be 2^N judgment modes. In this section, $N = 5$. For a joint distribution $p(x_1, x_2, x_3, x_4, x_5)$, the corresponding entropy $H(P(\{x_i\}))$ is defined as

$$H(p(\{x_i\})) = -\sum_{\{x_i\}} p(\{x_i\}) \log\, p(\{x_i\}). \tag{1}$$

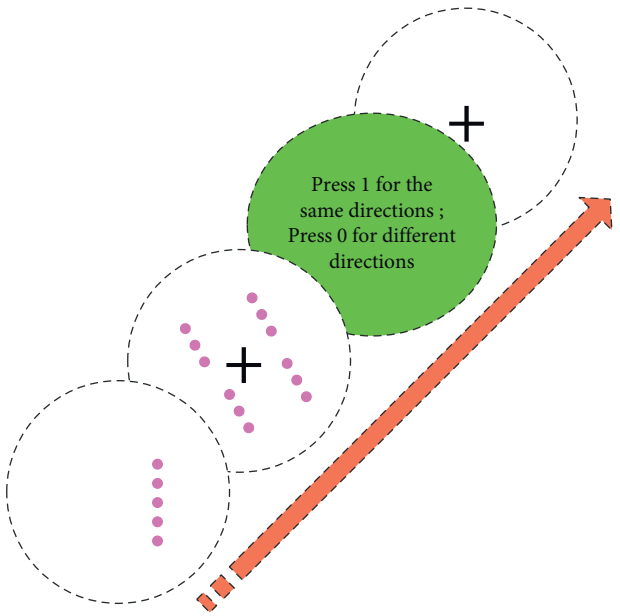

Figure 1: Schematic diagram of the language judgment test.

Among them, $\{x_i\}$ represents all possible judgment modes. The purpose of the principle of maximum entropy is to find a distribution P^* that maximizes its entropy under the constraints of experimental data.

$$p^*(\{x_i\}) = \arg\max_{p(\{x_i\})\in P} H(p(\{x_i\})). \tag{2}$$

Among them, p is constrained by the experimental data, which specifically refers to the constraints of various order correlations between the data. For example, for a second-order maximum entropy model,

$$\begin{aligned} L[p(\{x_i\})] = &-\sum_{\{x_i\}} p(\{x_i\})\log_2 p(\{x_i\}) \\ &+ \sum_i \alpha_i\left[\langle x_i\rangle_p - \langle x_i\rangle_{\text{data}}\right] \\ &+ \sum_{i<j} \beta_{ij}\left[\langle x_i x_j\rangle_p - \langle x_i x_j\rangle_{\text{data}}\right] \\ &+ \sum_{\{x_i\}} \lambda[p(\{x_i\}) - 1]. \end{aligned} \tag{3}$$

$\langle\cdot\rangle$ represents the data or the expectation under the maximum entropy model, and the Lagrange multiplier λ ensures the normalization of this distribution. For parameters α_i and β_{ij}, this is an optimization problem with a unique solution due to the convexity of entropy. The explicit expression after solving is as follows:

$$p^{(2)}(\{x_i\}) = \frac{1}{Z_2}\exp\left(\sum_i^N \alpha_i x_i + \sum_{i<j}^N \beta_{ij} x_i x_j\right). \tag{4}$$

The parameters α_i and β_{ij} here represent that the expectation of the data and the model are equal, and the second-order correlation between the data and the model is equal. Z_2 is the partition function, or normalization factor, to ensure that $p^{(2)}(\{x_i\})$ is a probability distribution and is expressed as follows:

$$Z_2 = \sum_{\{x_i\}} \exp\left(\sum_i \alpha_i x_i + \sum_{i<j} \beta_{ij} x_i x_j\right). \tag{5}$$

Similarly, the n-order maximum entropy model is as follows:

$$p^{(n)}(\{x_i\}) = \frac{1}{Z_n}\exp\left(\sum_i^N \alpha_i x_i + \sum_{i<j}^N \beta_{ij} x_i x_j + \cdots + \sum_{i<j<\cdots<n}^N \gamma_{12\cdots n} x_1 x_2 \cdots x_n\right). \tag{6}$$

Among them,

$$Z_n = \exp\left(\sum_i^N \alpha_i x_i + \sum_{i<j}^N \beta_{ij} x_i x_j + \cdots + \sum_{i<j<\cdots<n}^N \gamma_{12\cdots n} x_1 x_2 \cdots x_n\right). \tag{7}$$

In particular, if the assumptions are made to be independent of each other, then it is equivalent to a first-order model as follows:

$$p^{(1)}(\{x_i\}) = \frac{1}{Z}\exp\left(\sum_i^N \alpha_i x_i\right). \tag{8}$$

A core problem of the maximum entropy model is to solve the parameters from the experimental data. In this section, we use the Generalized Iterative Scaling algorithm to solve as follows:

$$\alpha_i^{\text{new}} = \alpha_i^{\text{old}} + c\cdot \log\left(\frac{\langle x_i\rangle_{\text{data}}}{\langle x_i\rangle_p}\right), \tag{9}$$

$$\beta_{ij}^{\text{new}} = \beta_{ij}^{\text{old}} + c\cdot \log\left(\frac{\langle x_i x_j\rangle_{\text{data}}}{\langle x_i x_j\rangle_p}\right), \tag{10}$$

$$\gamma_{12\cdots n}^{\text{new}} = \gamma_{12\cdots n}^{\text{new}} + c\cdot \log\left(\frac{\langle x_1 x_2 \cdots x_n\rangle_{\text{data}}}{\langle x_1 x_2 \cdots x_n\rangle_p}\right). \tag{11}$$

Among them, $c \in (0, 1)$ is used to control the convergence speed of the algorithm.

We used PsychoPy software to design linguistic experiments with 6 subjects, 3 males and 3 females. They are asked to judge whether a set of lines that appeared on the monitor were oriented the same or different, and each subject repeats the test 1,000 times. Specifically, we set x_1 to denote the subject's judgment as follows:

$$x_1 = \begin{cases} 1, & \text{identical}, \\ 0, & \text{different}. \end{cases} \tag{12}$$

x_2 is the number of lines as follows:

$$x_2 = \begin{cases} 1, & \text{the quantity is 4}, \\ 0, & \text{the quantity is 2}. \end{cases} \tag{13}$$

x_3 is the width of the line as follows:

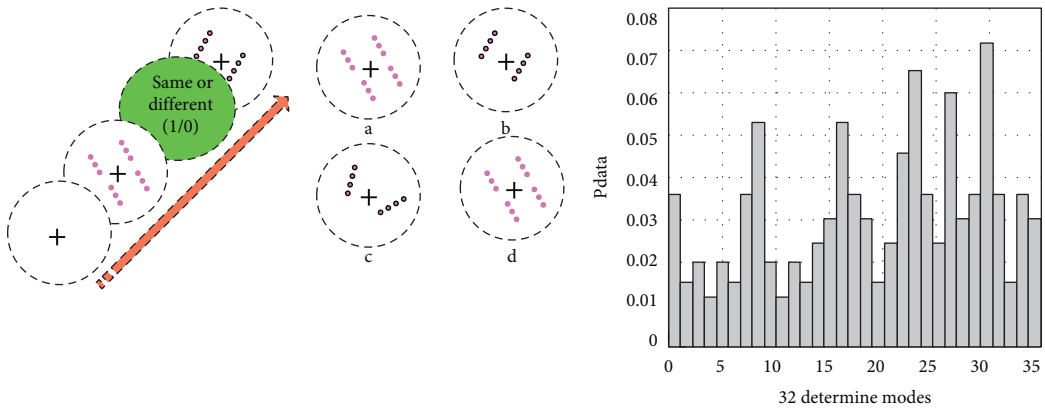

Figure 2: Test process and test data probability.

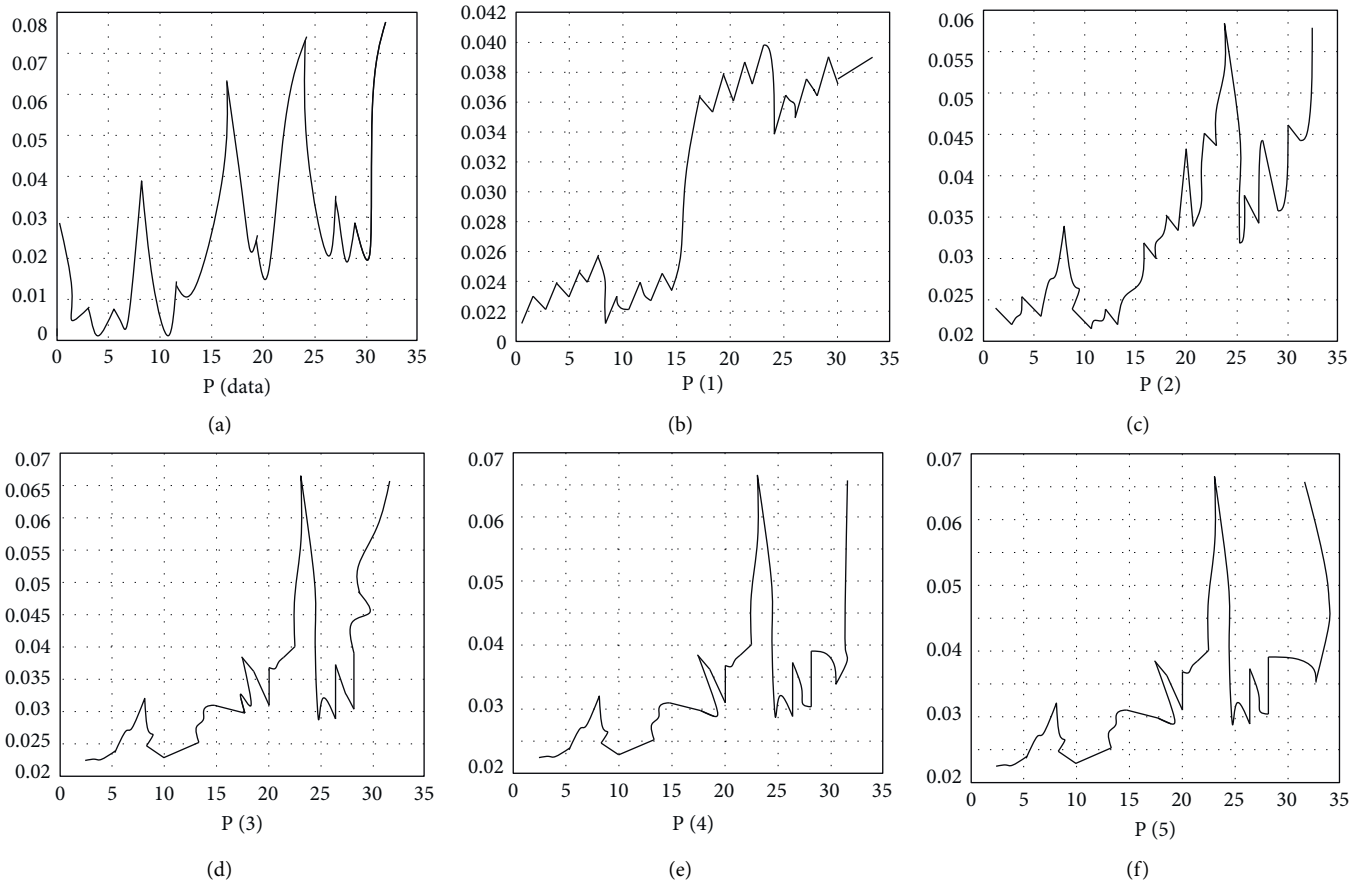

Figure 3: Data and probability values of different language judgment models.

$$x_3 = \begin{cases} 1, & \text{width 5,} \\ 0, & \text{width 1.} \end{cases} \tag{14}$$

x_4 is the length of the line as follows:

$$x_4 = \begin{cases} 1, & \text{length 0.5,} \\ 0, & \text{width 0.2.} \end{cases} \tag{15}$$

x_5 is the color of the line as follows:

$$x_5 = \begin{cases} 1, & \text{the color is red,} \\ 0, & \text{the color is gray.} \end{cases} \tag{16}$$

The vector $\vec{x} = (x_1, x_2, x_3, x_4, x_5)$ is called a judgment mode.

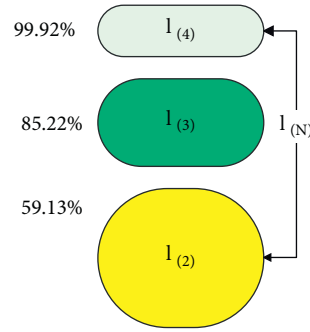

Figure 4: Contribution of each order model.

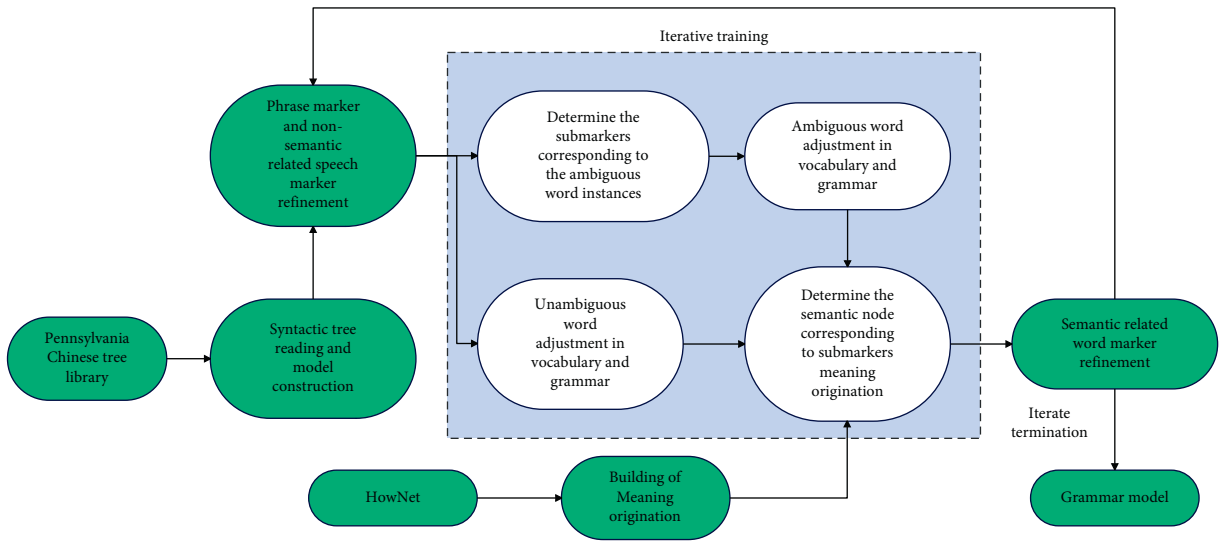

Figure 5: Block diagram of word sense disambiguation in the training process of English syntactic analysis.

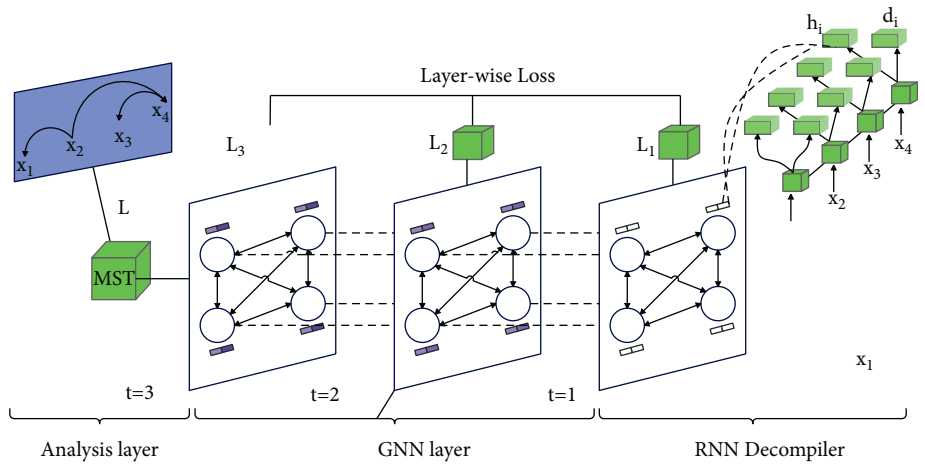

Figure 6: Dependency English syntax analyzer based on graph neural network.

Figure 2(a) represents the sequence in which the experiments were performed. First, there will be a cross on the screen to keep the subject's attention in the middle of the screen. After that, a stimulus pattern is randomly generated, and the subject needs to make a judgment as to whether they are oriented in the same direction. If the direction is the same, they press "1," and if the direction is different, they press "0." After that, it will enter the next judgment process, and in this cycle, each subject will perform 1000 times. Figure 2(b) randomly enumerates 4 different stimulation

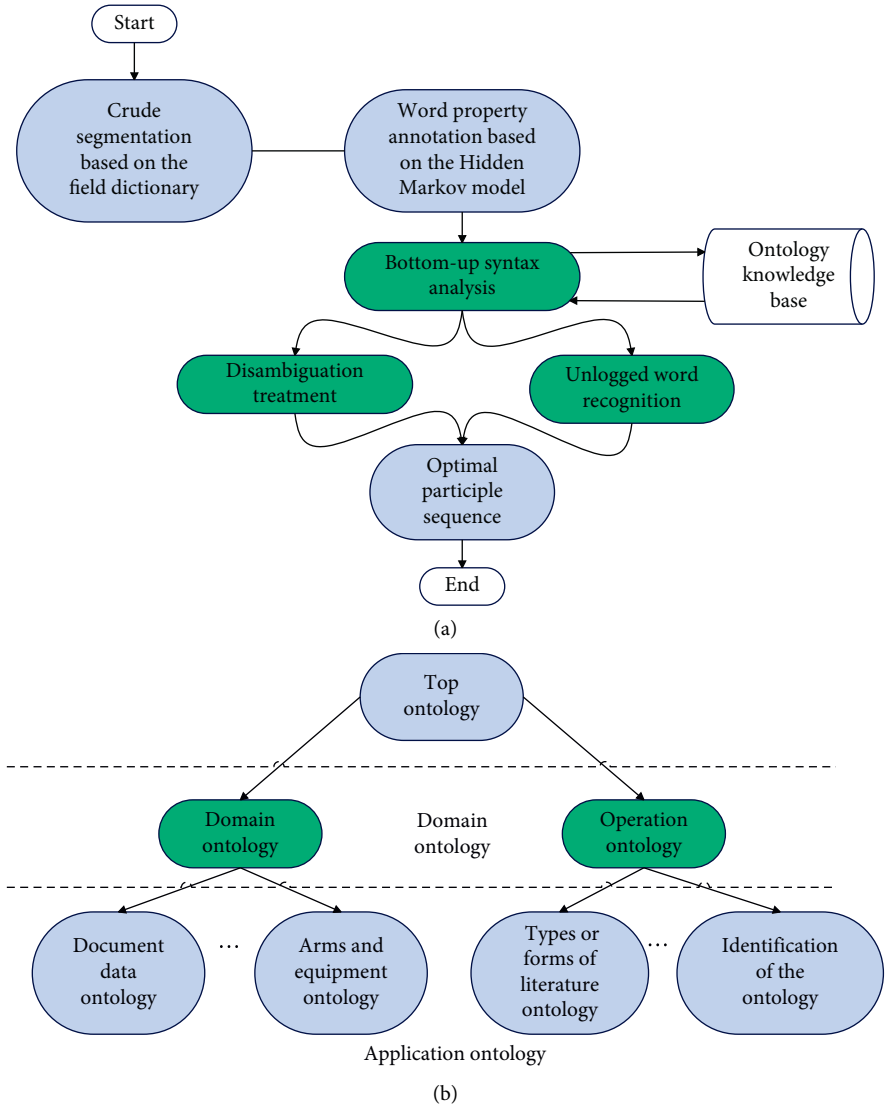

Figure 7: Word segmentation level model. (a) Word segmentation framework. (b) Ontology level.

patterns, and the number, color, length, and width of the lines vary. Figure 2(c) is a histogram of the test results of the subjects, the abscissa represents different stimulation patterns, and the ordinate is the corresponding probability.

It can also be seen from Figure 2(c) that for different stimuli, the subjects' judgments are quite different. For example, subjects in the mode (1,1,1,1) were twice as likely to judge that the lines were parallel as in the mode (0,0,0,0). Therefore, some stimulation conditions can promote the correct rate of cognition, and different stimulation modes also have different effects. We wanted to know what degree of correlation had a significant impact.

We use experimental data to train the model and solve for the parameters. After that, we calculate the first-order model $p^{(1)}$ and the second-order model separately $p^{(2)}$, until the end of the fifth-order model $p^{(5)}$.

For a distribution $p(\{x_i\})$, the standard formula for its entropy H is

$$H = -\sum_{\{x_i\}} p(\{x_i\}) \log\, p(\{x_i\}). \tag{17}$$

It should be noted here that the entropy H_1 of the first-order model is always larger than the entropy $H_2, H_3, \ldots, H_N$ of the higher-order model because of the entropy difference between the phase and the experimental data:

$$I_N = H_1 - H_N. \tag{18}$$

The contribution of the K-order model is given by $I_{(K)}$ as follows:

$$I_{(K)} = H_{K-1} - H_K. \tag{19}$$

Thus, the expressivity of the K-order model can be represented by the ratio r_K as follows:

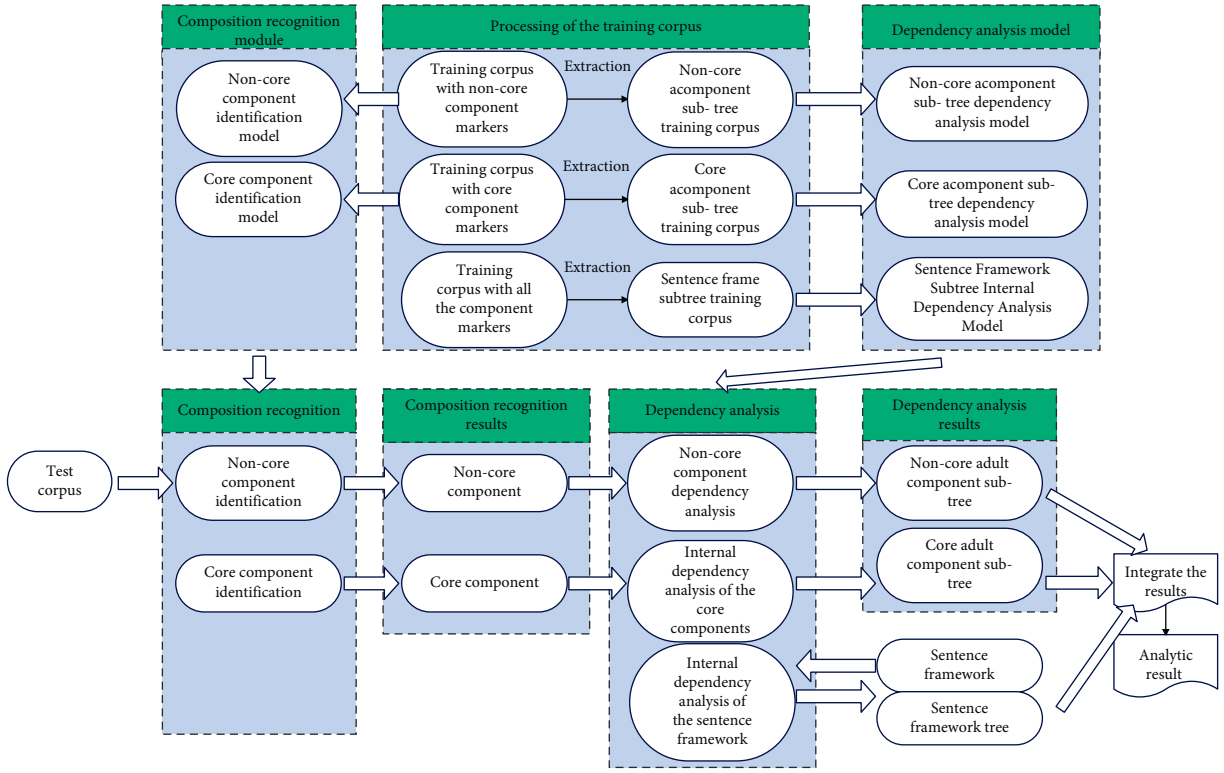

Figure 8: Test process diagram.

Table 1: Syntactic analysis effect.

Num	Syntax analysis	Num	Syntax analysis	Num	Syntax analysis
1	85.83	16	82.00	31	85.70
2	78.71	17	72.24	32	72.57
3	80.69	18	85.67	33	80.83
4	78.52	19	76.33	34	82.65
5	77.91	20	80.06	35	69.85
6	73.37	21	75.74	36	75.96
7	76.96	22	81.08	37	75.66
8	84.70	23	68.64	38	77.36
9	71.53	24	74.09	39	85.39
10	84.44	25	85.19	40	83.98
11	78.88	26	84.67	41	70.38
12	69.13	27	73.15	42	84.29
13	68.66	28	75.29	43	78.38
14	72.03	29	76.88	44	78.98
15	75.06	30	73.41	45	75.20

$$r_K = \frac{I_{(K)}}{I_N}. \tag{20}$$

In Figure 3, the probability $p^{(\text{data})}$ of the test data and the first-order model $p^{(1)}$, the second-order model $p^{(2)}$, the third-order model $p^{(3)}$, the fourth-order model $p^{(4)}$, and the fifth-order model $p^{(5)}$ is shown in turn. The abscissa is the different judgment modes, there are $2^5 = 32$ kinds in total, and the ordinate is the test data probability value or model prediction value of the corresponding judgment mode. As can be seen from the figure, there is a large gap between the first-order model and the experimental data, but among the remaining models, we cannot judge which one is better, so we need to use entropy to quantify it.

The proportional score r_K is calculated using the formula (20), and the result is shown in Figure 4. Among them, $I_N = I_{(2)} + I_{(3)} + I_{(4)} + I_{(5)}$. After that, the contribution of the second- to fifth-order models to the correlation can be calculated. It is found by calculation that the third-order model contains 85.22% of the information in I_N. Therefore, in the language judgment test, the relevant data needs to consider higher-order correlations, and only considering the second-order correlations is not enough to accurately describe the test data and the independence assumption used to simplify the model is more deviated from the actual situation.

3. English Syntactic Analysis and Word Sense Disambiguation Strategy of Neutral Set from the Perspective of Natural Language Processing

The English syntactic analysis training and word sense disambiguation process redesigned in this paper are shown in Figure 5. The dotted box is the part that implements word sense disambiguation, including the processing of unambiguous words and the processing of ambiguous words. In the unambiguous word processing part, the system firstly adjusts the lexical grammar model obtained from the previous round of training, so that a certain word only

Table 2: Word sense disambiguation effect.

Num	Disambiguation effect	Num	Disambiguation effect	Num	Disambiguation effect
1	74.88	16	71.86	31	79.38
2	79.12	17	66.68	32	65.81
3	74.97	18	70.56	33	75.08
4	78.91	19	74.62	34	72.80
5	72.16	20	72.17	35	65.44
6	81.83	21	72.46	36	74.59
7	79.72	22	67.63	37	67.44
8	82.24	23	79.80	38	82.70
9	80.48	24	75.29	39	80.02
10	79.01	25	78.62	40	82.66
11	67.03	26	69.75	41	73.87
12	82.89	27	77.98	42	72.77
13	78.50	28	69.39	43	70.06
14	74.88	29	68.77	44	74.60
15	68.41	30	78.89	45	67.83

corresponds to the most likely part-of-speech subtag. That is, through adjustment, the system makes the set of words corresponding to a certain part-of-speech subtag different from other part-of-speech subtags. In the ambiguous word processing part, since disambiguation can only be done by returning to the instance, this part first traverses all the English syntax trees, examines each polysemy, and determines its most likely corresponding part-of-speech tag according to the English syntax information. Next, the system modifies the lexical grammar and puts the polysemy into the vocabulary set of the corresponding part-of-speech subtag.

This paper uses BiLSTM to vector-encode words and then uses first-order dependency information to build a graph structure between words. Moreover, this paper performs graph neural network computations on this graph structure, iteratively updating the vector representation of each word. The update formula is designed to introduce information on second-order dependency structures (including possible grandfathers, grandchildren, and brothers). After several iterative updates, the vector representation of each word is used to compute the score for each possible dependent edge. Finally, the classical first-order dependency analysis algorithm is used to decode the dependent English syntax tree. Figure 6 shows the neural network architecture of the method.

As shown in Figure 7(a), the processing process of the compound word segmentation method is as follows: after obtaining the document to be segmented, the system performs preliminary processing on the text, including word segmentation based on dictionary and part-of-speech tagging based on the hidden Markov model. Then, the system performs bottom-up English syntactic analysis on the processed sentences and finally performs disambiguation processing and identification of unregistered words to obtain the optimal word segmentation sequence.

The ontology definition proposed in this paper is multilevel, which can better realize the sharing and reuse of the ontology, and its level is shown in Figure 7(b). Domain ontology and subontologies of operation ontology are contained relationships. Moreover, domain ontology focuses on the establishment of terms, which is aimed at the definition of object properties in the domain. In the operation ontology, the genre ontology is the definition of the grammar and words of a specific genre, and the recognition ontology is the definition of the rules used in the recognition of unregistered names.

In order to simplify the sentence structure to eliminate long-distance dependencies, this paper proposes a dependent English syntax analysis that introduces the hierarchical component analysis method. The specific experimental process is shown in Figure 8.

For the training corpus, three kinds of the corpus are obtained by using the component subtree labeling algorithm: the training corpus marked with noncore components, the corpus marked with core components, and the corpus marked with all components. Among them, the training corpus marked with noncore components is used to train the noncore component recognition model and extract the noncore component subtrees. The training corpus marked with core components is used to train the core component recognition model and extract the core component subtree, and the corpus marked with all components is used to generate the sentence frame subtree. Finally, the three seed tree corpora are used to train their internal dependency analysis model.

Among them, the model of the noncore component recognition system uses the Chinese prepositional phrase recognition system based on the cascading conditional random field, and the model of the core component recognition system uses the model generated from the training corpus marked with the core component.

On the basis of the above research, the effectiveness of the English syntactic analysis and sense disambiguation strategy of neutral set words proposed in this paper is verified from the perspective of natural language processing. This paper counts the syntactic analysis effect and word sense disambiguation effect of the system model in this paper, and the experiment in this paper is carried out on the MATLAB platform, and the statistical test results are shown in Tables 1 and 2.

From the above simulation experiments, it can be seen that the English syntactic analysis and word sense

disambiguation strategy of neutral set words proposed in this paper from the perspective of natural language processing is very effective.

4. Conclusion

Since natural language processing is directly application-oriented, it requires that there must be a suitable method to process languages in batches and accurately find the required information, and all subsequent operations are based on this basis. In concrete practice, dependency grammar has proved to be a suitable theory, and scholars have established relatively advanced syntactic analysis methods based on the basic idea of dependency grammar. It can be said that the theory and analysis methods of dependency grammar have provided great help to ontology research and applied research of linguistics. With the support of natural language processing ideas, this paper analyzes English syntactic analysis and word sense strategy of the neutral set word. The simulation experiment study shows that the English syntactic analysis and word sense disambiguation strategy of neutral set words proposed in this paper is very effective from the perspective of natural language processing.

Data Availability

The labeled dataset used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was sponsored by Zhengzhou Railway Vocational and Technical College.

References

[1] C. Wang, Y. Zhao, and D. Sun, "Research on design and sharing of yi language corpus Resources Database based on syntactic rules," *Solid State Technology*, vol. 63, no. 5, pp. 10563–10574, 2020.

[2] M. Snaith, N. Conway, T. Beinema et al., "A multimodal corpus of simulated consultations between a patient and multiple healthcare professionals," *Language Resources and Evaluation*, vol. 55, no. 4, pp. 1077–1092, 2021.

[3] M. Esplà-Gomis and A. Sentí, "Presentació del monogràfic «Spoken Corpus Linguistics in Romance: thoughts, design and results," *Caplletra. Revista Internacional de Filologia*, vol. 69, pp. 117–123, 2012.

[4] A. A. A. Alkhalifa and I. B. E. Mohammed, "Corpus-based, genre-Analytic Approach to Discipline-specific materials design and development," *Bulletin of Advanced English Studies–*, vol. 3, no. 1, pp. 34–43, 2019.

[5] D. Knight, F. Loizides, S. Neale, L. Anthony, and I. Spasić, "Developing computational infrastructure for the CorCenCC corpus: the National corpus of contemporary Welsh," *Language Resources and Evaluation*, vol. 55, no. 3, pp. 789–816, 2021.

[6] S. Granger and M. A. Lefer, "The multilingual student translation corpus: a resource for translation teaching and research," *Language Resources and Evaluation*, vol. 54, no. 4, pp. 1183–1199, 2020.

[7] P. De Graaf, R. Ramadan, E. C. Linssen et al., "The multilayered structure of the human corpus spongiosum," *Histology & Histopathology*, vol. 33, no. 12, pp. 1335–1345, 2018.

[8] G. Santos, "Designing and building SCoPE^2: a spoken corpus of Brazilian Portuguese and L2-English," *Research in Corpus Linguistics*, vol. 8, no. 1, pp. 49–64, 2020.

[9] M. R. Peñarroja, "Corpus Pragmatics and Multimodality: Compiling an Ad-Hoc multimodal corpus for EFL Pragmatics teaching," *International Journal of Instruction*, vol. 14, no. 1, pp. 927–946, 2021.

[10] M. d. Mar Sánchez Ramos, "Documentation in specialised contexts: a quasi-experimental corpus-based study in public service interpreting and translation studies," *Translation and Translanguaging in Multilingual Contexts*, vol. 8, no. 1, pp. 30–48, 2022.

[11] D. Gablasova, V. Brezina, and T. McEnery, "The trinity Lancaster Corpus: development, description and application," *International Journal of Learner Corpus Research*, vol. 5, no. 2, pp. 126–158, 2022.

[12] D. Zeyrek, A. Mendes, Y. Grishina, M. Kurfalı, S. Gibbon, and M. Ogrodniczuk, "TED Multilingual Discourse Bank (TED-MDB): a parallel corpus annotated in the PDTB style," *Language Resources and Evaluation*, vol. 54, no. 2, pp. 587–613, 2020.

[13] N. Smith and C. Waters, "From broadcast archive to language corpus: Designing and investigating a sociohistorical corpus from Desert Island Discs," *ICAME Journal*, vol. 42, no. 1, pp. 167–190, 2018.

[14] P. D. Loprinzi, J. Harper, and T. Ikuta, "The effects of aerobic exercise on corpus callosum integrity: systematic review," *The Physician and Sportsmedicine*, vol. 48, no. 4, pp. 400–406, 2020.

[15] R. F. Alfuraih, "The undergraduate learner translator corpus: a new resource for translation studies and computational linguistics," *Language Resources and Evaluation*, vol. 54, no. 3, pp. 801–830, 2020.

[16] J. Tanusy, "Men and Women in Suicide Notes: a corpus-based Rhetorical Moves analysis," *Journal of Language and Literature*, vol. 22, no. 1, pp. 64–74, 2022.

[17] X. Yang, "A corpus-based study of modal Verbs in Chinese Learners' Academic Writing," *English Language Teaching*, vol. 11, no. 2, p. 122, 2018.

[18] S. Li and C. Kit, "Legislative discourse of digital governance: a corpus-driven comparative study of laws in the European Union and China," *International Journal of Legal Discourse*, vol. 6, no. 2, pp. 349–379, 2021.

[19] N. Herry-Bénit, S. Lopez, T. Kamiyama, and J. Tennant, "The interphonology of contemporary English corpus (IPCE-IPAC)," *International Journal of Learner Corpus Research*, vol. 7, no. 2, pp. 275–289, 2021.

[20] A. Matthews, "Sociotechnical imaginaries in the present and future university: a corpus-assisted discourse analysis of UK higher education texts," *Learning, Media and Technology*, vol. 46, no. 2, pp. 204–217, 2021.