Feature recognition of English clauses based on particle swarm optimization algorithm

Liu Aiqin

Shandong Vocational College of Light Industry, Zibo, 255300, China

Keywords: Particle swarm optimization; English clauses; Feature recognition

Abstract: Feature recognition of English clauses is a basic problem of syntactic analysis. It is the basis of English-Chinese machine translation. A feature recognition method of English clauses based on particle swarm optimization algorithm is proposed. This paper analyzes the characteristics of English clauses, delimits the boundary of clauses, and follows the current optimal particle in the solution space to search the best position through the cooperation and information sharing between particle swarm individuals. The feature set is selected, the crossover and mutation idea of genetic algorithm is introduced, and the crossover operation is carried out to complete the feature recognition of English clauses. The experimental results show that when the threshold P is 50, the recognition accuracy of this algorithm is consistent with that when p is 100, and the recognition accuracy is 93.45%. The accuracy of particle swarm optimization algorithm for English clause feature recognition is high, which remains at about 90%. Compared with the two literature methods, the convergence performance of particle swarm optimization algorithm is better.

1. Introduction

A clause is a grammatical unit that contains at least one subject and predicate and expresses a point of view [1]. Clause recognition refers to the process of marking the level of clauses according to their grammatical structure. It belongs to the category of shallow syntactic analysis. Shallow layer the main task of syntactic analysis is the recognition and analysis of chunks, which simplifies the task of syntactic analysis to some extent.

It is also the basis for further analysis of sentences [2]. In natural language processing, syntactic analysis research mostly focuses on single sentences and has achieved great success [3-4]. Consider simplifying complex sentences before parsing them. For sentences with clauses. This simplified process can be realized by identifying English clauses first. Then analyze the subject sentence and clause respectively. The purpose of clause recognition is to reduce the complexity of in-depth analysis of sentences by defining the boundary of clauses.

Psycholinguists have found that when reading a sentence, people always find the boundary of the clause before dealing with the long sentence, and cut the long sentence into several sub sentence blocks for understanding respectively. This inspires researchers whether they can first divide the sentences of the source language into several sub sentences and simplify the sentence pattern of the source sentence in natural language processing, which is conducive to in-depth analysis. For the recognition of English clauses, researchers at home and abroad have put forward some research

methods. Some researchers have proposed using machine learning algorithm to determine the boundary of clauses in text [5]. Some researchers use hidden Markov model to identify the boundary of clauses. Some researchers use the connectives of clauses and other functional words to establish a dictionary or grammar rule library to segment long sentences; or based on corpus, using part of speech information, combining rules and statistics to identify the boundary of clauses.

From the perspective of machine learning, literature [6] introduces multiple features to improve the lack of information caused by Chinese small data sets. In addition, this study decomposes the character recognition problem into question and answer text matching problem, answer text implication problem and standard answer text implication problem, and continues to train on the short text score data set. The final result has some improvement, which proves the usability of the mechanism designed in this paper. Literature [7] proposed an ancient recognition neural network method based on multi-level convolution to help artists imagine the appearance of this ancient painting after restoration. If analysis is introduced to match the map by adding missing regions and nearest neighbor pixels to enhance the rough estimation of the complete image. Domain specific pyramid networks are used to capture various spatial context quantities. Reference [8] proposed the introduction of TM with integer weighting clause, i.e. integer weighting TM (IWTM), which solves the accuracy interpretation challenge in machine learning. The purpose is to improve the interpretability of TM by reducing the number of terms required for competitive performance. IWTM achieves this through weighted clauses so that a single clause can replace multiple repeated clauses. Since each TM clause is adaptively formed by tsetlin automata (TA) team, identifying effective weights has become a challenging online learning problem. This problem is solved by using another kind of automata to extend each TA team: Online random search (SSL) automata.

In this paper, an English clause feature recognition method based on the particle swarm optimization algorithm is proposed, which can effectively increase the diversity of the population by crossover and mutation of particles.

2. An analysis of the characteristics of English clauses

In terms of composition, compound sentences can be divided into the following two types:

(1) Connect the clause with the main sentence with connectives (in some cases, connectives can be omitted).

If you're not good at figures, (clause) it is pointless to apply for a job in a bank (main sentence).

(2) Use verb infinitives or participle structures, which form part of a compound sentence rather than a simple sentence.

To get into university you have to pass a number of examinations.

Here, the second case is treated as a general phrase, mainly considering the recognition of clauses in the first case.

The identification of English clauses in this paper is mainly the delimitation of the boundary of clauses [9].

This paper is based on Penn tree bank corpus, which is an English corpus with various levels of annotation, such as part of speech level, phrase level, clause level and so on. 21115 clauses are extracted, including 4936 recursive clauses. By analyzing the structure of these sentences, it is found that the composition of clauses, whether recursive or non-recursive, can be divided into the following three cases:

(1) Guided by certain functional words, these functional words include who, which, when and other wh words, conjunctions such as before, after, if and so on:

(2) Special verbs are followed by object clauses, such as say, think, etc. many adjectives describing personal feelings (such as afraid, glad, etc.) or adjectives expressing certainty (such as

certain, sure, etc.) are followed by clauses [10];

(3) Special sentence structure, if a predicate verb is preceded by two consecutive BNPs (basic noun phrases), the predicate verb is most likely a clause verb, or if the two predicate verbs are together, one of them must belong to a clause verb.

Statistics show that the first category accounts for a large proportion, about 75%, and the third category of clauses accounts for a small proportion, only 710. According to the statistics of the training corpus, the priority of these three clear situations is decreasing. At the same time, in each case, they are arranged according to the probability from large to small. For example, the priority of wh word is higher than that of other conjunctions. In this way, in a complex sentence, if the above situations occur comprehensively, it can be determined according to the priority [11]. In the recognition of the beginning of a clause, it is processed separately according to the different situations of the clause. Through the analysis of the clause in the corpus, it can be concluded that the position of the end of the clause is related to the following factors: the position of the clause in the sentence, the relationship between the clause and the main sentence, the position of the recognizing the end of a sentence, as well as some punctuation marks, and so on. When recognizing the end of a sentence, process it according to the above information.

3. Feature recognition of English clauses based on particle swarm optimization

3.1 Basic principle of particle swarm optimization algorithm

When each element in Y is given an optimized weight[12], we can not only retain the number of features we need, but also highlight the role of useful elements in the features. Therefore, the particle dimension is equal to the feature dimension. Randomly initialize each particle of particle swarm Z (particle swarm size N), set the initial position of the i particle as $Z_i = [Z_{i1}, Z_{i2}, \dots, Z_{im}] (i \le N)$ and the particle dimension т as . The speed is $V_i = [V_{i1}, V_{i2}, ..., V_{im}], 1 \le i \le N$. In order to prevent the target from crossing the boundary, premature or falling into local minimization, the position and velocity range of particles are limited $-0.5 \le Z_{ij} \le 0.5, 1 \le V_{ij} \le 1$. The fitness function of the *i* particle is given by the following formula:

$$F(i) = \frac{1}{N} \sum_{j=1}^{N} -e^{\frac{y_{ij}^2}{2}}$$
(1)

Among them

$$y_i = Z_i Y = [y_{i1}, y_{i2}, ..., y_{iN}], (1 \times N)$$
(2)

Equation (1) is used as the fitness function because when each feature is multiplied by the corresponding optimization coefficient, the position with the smallest function value is the most optimal position [13]. At this time, the correlation between each one-dimensional feature is the smallest and the discrimination between classes is the largest. At the same time, the function is a Gaussian function, which can reduce the absolute value of the result and improve the recognition efficiency.

For the i particle, after calculating the fitness function value in the t cycle, compare its fitness function value with its original historical optimal position Pbest(i). If it is better, update Pbest(i) with the latest $Z_i(t)$, and then compare it with the global optimal position Gbest. If it is better,

take it as the current Gbest. After each comparison, use equations (3) and (4) to update the position and speed.

$$v_{ij}(t+1) = wv_{ij}(t) + \eta_1 r_1 (Pbest - x_{ij}(t)) + \eta_2 r_2 (Gbest - x_{ij}(t))$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1)$$
(3)

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1)$$
(4)

 $v_{ij}(t+1)$ indicates that the velocity of the *i* particle is the ^j dimensional element in the t+1 iteration, and $x_{ij}(t+1)$ indicates that the position of the *i* particle is the ^j dimensional element in the t+1 iteration; *w* is called inertia weight, $\eta_1\eta_2$ is two normal numbers, called learning factor, and r_1, r_2 is two uniform random numbers between [0,1], [14].

When the iteration termination conditions are met, the iteration process is ended. Experiments show that the number of iterations of termination condition selection reaches the predetermined number, and the recognition result is better. The final *Gbest* is the optimal solution $Z^* = [Z_1^*, Z_2^*, ..., Z_m^*]$

3.2 Feature constraints of clause recognition

In Clause feature recognition, the selection of feature set is a very important problem. The following will introduce the features used in Clause recognition.

Generally speaking, the features of sentence beginning recognition are divided into two categories: lexical features and sentence features. The lexical features adopt the sliding window method to obtain three types of features: word, part of speech tagging and phrase tagging. Second, sentence features can be divided into five aspects: sentence structure, function word information, verb information, punctuation information and special circumstances.

(1) Sentence structure

① Whether the current position is the beginning of a sentence is a feature;

⁽²⁾ The part of speech string and phrase string of the left and right parts of the sentence are extracted, and only verb phrases, commas and related words are concerned in the phrase string [15-16].

(2) Functional word information

① When the current word is if / that / what / who / where / when / why / who / how / while, the feature is obtained by labeling the category with the word;

(2) When the current word is which, check whether its previous word is at / in / on, etc., and get the feature combined with its annotation category.

(3) Verb information

Taking the current word as the boundary, take the number of VPS in the left and right parts of the sentence (including the case where the number is zero) as a feature.

(4) Punctuation information

① If there is only one comma in the sentence, it can be divided into the whole sentence; If there are multiple commas in the whole sentence, take whether there is VP between commas as a feature information [17-18];

2 When the current word is a colon or quotation mark, take the annotation of the word itself and the following word as a feature.

(5) Special circumstances

① When the current word is and or or, take whether there is VP around it as a feature;

⁽²⁾ When the prototype of the current word is say, the word itself constitutes a feature with its annotation category.

The features used for sentence beginning and sentence end recognition are basically the same, only the latter also uses the estimation results of the former as a class of features.

For the annotation of complete clauses, this paper mainly follows Xavier's idea. First, judge whether the predicted clause head position is multiple clauses, then find out the possible clause candidate set for the starting position of each possible clause, and finally get the most appropriate clause annotation from the scoring function. In this part, in addition to the previously used features, the prediction results of the previous two steps are also introduced as known information. The phrase string features from the prediction start boundary to each prediction end position are added to the sentence features.

After feature selection, they need to be coded [19-20], and the features are represented by binary functions in Clause feature recognition:

$$f_{j}(a,b) = \begin{cases} 1 & \text{if} a = S \text{ and } h(b) = 1 \\ 0 & \text{else} \end{cases} \quad 1 \le j \le k$$
(5)

Here, *a* represents category, *b* represents environment information, and function h(b) represents predicate. A feature consists of category and predicate. The value of the predicate is 1/0. In use, only those features whose value is equal to 1 are selected, that is, those features that only appear in training.

3.3 Recognition of English clause features by particle swarm optimization

In order to overcome the problem that the population can easily fall into local extremum, the crossover and mutation ideas of genetic algorithm are introduced. Before the end of each iteration, the particles in the population cross operate with a probability of 0.5 to generate a new generation of particles. After the crossover operation is completed, select those particles whose fitness value is worse than the average fitness value of the population [21-22], and mutate them. The variation process is as follows:

$$m_r = \begin{cases} -x_i, & rand < p_m \\ x_i, & otherwise \end{cases}$$
(6)

Where P_m is the variation probability. *rand* is a random number uniformly distributed on [0,1]. In the late iteration of the algorithm, because the diversity of the population is enriched, the population can be prevented from falling into local extremum.

The feature recognition of English clauses based on particle swarm optimization algorithm is as follows:

1) Initialization of particles: all N training samples are numbered, and each training sample corresponds to a number in 1-N and does not repeat each other [23]. Let the number of particles be l, take one particle randomly from N training samples, take the number of this training sample as the English clause feature $x_l (l = 1, 2, ..., L)$ of the particle, and initialize the particle speed to be equal to 0

2) Initialize particle fitness value: for particle l, according to the English clause feature x_l of

the particle, the sample corresponding to particle l can be uniquely found in N training samples [24]. The correlation coefficient between the training sample and the test sample is recorded as d_i , then d_i can be used as the fitness value of particle l, that is:

$$f(x_i) = d_1, l = 1, 2, \dots, L \tag{7}$$

Repeat step 2 for all particles to obtain the initial fitness value.

3) Initialize individual optimal English clause features and global optimal English clause features: the initial individual optimal English clause features of each particle are the initial English clause features of the particle, namely:

$$p_{ld} = x_l, l = 1, 2, ..., L$$
 (8)

The fitness value corresponding to the initial optimal English clause feature is the initial fitness value of the particle. For the particle with the largest initial fitness value, the English clause is characterized by the initial global optimal solution;

4) Update the speed of each particle and the characteristics of English clauses ;

5) Update the fitness value of each particle: method synchronization step 2;

6) Update the individual optimal English clause features and global optimal English clause features: for particle l, Set $x_l(t+1)$ as the updated English clause features, then update the individual optimal English clause features and individual optimal fitness values.

7) The global optimal English clause feature and its fitness are worth updating. It is not only for a particle, but also needs to traverse the whole population.

8) Termination conditions: there are two main cases of iterative termination conditions: one is to find a good enough solution; Second, the loop reaches the maximum number of iteration steps. For the first case, this paper defines two termination conditions:

$$\mathbf{a.}\left|f(p_{gd})\right| \ge S \tag{9}$$

Among them, P_{gd} represents the global optimal fitness value, and S is usually set to a value of 0.0-1.00. At this time, it is considered that particle swarm optimization algorithm can search the optimal English clause characteristics of particles.

$$b. count(P_{gd}) = R \tag{10}$$

 $count(P_{gd})$ is an integer value indicating the number of times the global optimal particle of the whole population recognizes its English clause features in the iterative process. R is usually set to an integer value in the range of 15-30. When $count(P_{gd}) = -R$, it is considered that particle swarm optimization algorithm recognizes the optimal English clause features of particles.

8) Decision output: if the program does not meet the iteration termination conditions, return to step 4 to continue the iteration; If the program meets the iteration termination conditions, the training sample number with the largest correlation coefficient with the test sample is obtained. According to this number, the target category of the corresponding training sample can be found and determined as the test sample type.

So far, the recognition of English clause features based on particle swarm optimization has been completed.

4. Experiment

The experiment uses Penn tree bank corpus, in which wsj15-18 (211727 words) is used as the training set and wsj20 (47377 words) and wsj21 (40039 words) are used as the test set. Due to the uneven distribution of the two types of data in the training set, the characteristics of vocabulary becoming non sentence beginning (non sentence end) contribute much more than those becoming sentence beginning (sentence end). Therefore, in the experiment, instead of taking all words as the research object, the words at the phrase boundary are selected as candidates. Test the effect of particle swarm optimization algorithm in English clause feature recognition.

4.1 Experimental parameter setting

The experimental verification environment is amdathlon (th) $64 \times 23600 + CPU$, 1G memory, operating environment is Windows XP operating system, and verification platform is MATLAB 7.0. In this experiment, the number of particles is obtained according to the following formula:

$$popsize = round(12 + \sqrt{2*D}) \tag{11}$$

Where, D is the dimension of particles, that is, the number of features. The *round* function is used for rounding operations. The maximum number of iterations T_{max} is 100. The learning probability pc is set to 0.5. The coefficient β of iBQPSO algorithm is updated according to the following formula:

$$\beta = (0.95 - 0.55) * \frac{T_{\text{max}} - t}{T_{\text{max}}} + 0.55$$
(12)

Where, t is the current number of iterations. The experiment was conducted for 10 cycles. The optimal feature subset is obtained by particle swarm optimization algorithm.

4.2 Experimental results and analysis

4.2.1 Determination of filtering threshold

The setting of threshold P is very important because it determines the search space of particle swarm optimization algorithm. If the p value is set too large, the computational complexity of the algorithm will be increased, and the algorithm may not converge to the optimal solution. When the p value is set too small, the features that are helpful for recognition may be deleted, resulting in the reduction of the final recognition effect. We have conducted experiments on the three values of P taken as 50, 100, and 200 compared the algorithm in this paper with the methods in literature [6] and literature [7]. The experimental results are shown in Table 1:

Threshold P	Algorithm in this	Literature [6]	Literature [7]
	paper	methods	methods
50	93.45	80.16	86.48
100	93.45	83.45	87.36
200	92.64	79.79	85.97

Table 1: Comparison of recognition accuracy of different methods

It can be seen from table 1 that when the threshold P is 50, the recognition accuracy of the algorithm in this paper is consistent with that when p is 100, and the recognition accuracy is 93.45%. The recognition accuracy of the methods in literature [6] and literature [7] is worse than that in the case of P 100. When p is 200, the recognition accuracy is reduced in all three methods. To sum up, we select the first 100 features as the primary feature subset.

4.2.2 Comparison of algorithm optimization performance

In order to verify the effectiveness of particle swarm optimization algorithm in this paper, the above two literature methods are used as comparison methods, and the accuracy of this algorithm is verified through this group of comparison experiments. See Figure 1 for details.



Figure 1: Accuracy comparison

As can be seen from Figure 1, the accuracy rate of particle swarm optimization algorithm for English clause feature recognition is high, which remains at about 90%, while the accuracy rate of the two literature methods is less than 80%, and the accuracy rate of literature [7] algorithm is the highest, only reaching 75%, and the recognition effect is poor.

4.2.3 Analysis of convergence performance of algorithm

In order to verify the convergence performance advantages of particle swarm optimization algorithm, the convergence performance of the above two literature methods are compared. See Figure 2 for details.



Figure 2: Convergence of different methods

As can be seen from Figure 2, the particle swarm optimization algorithm converges to the highest recognition accuracy after about 40 iterations. The two literature methods need more than 60 iterations to converge to their highest recognition accuracy. Compared with the two literature methods, particle swarm optimization algorithm has better convergence performance. It can be seen that the recognition effect of particle swarm optimization algorithm is better. This is mainly because particle swarm optimization algorithm introduces the crossover and mutation idea of genetic algorithm in order to overcome the problem that the population is easy to fall into local extreme points. To ensure that the recognition rate of English clause features is effectively improved, and then improve its effectiveness.

5. Conclusion

This chapter proposes a feature recognition method of English clauses based on particle swarm optimization algorithm. The identification of English clauses in this paper is mainly the delimitation of the boundary of clauses. In the recognition of the beginning of a clause, it is processed separately according to the different situations of the clause. The method of sliding window is used to get three kinds of features: word, part of speech tagging and phrase tagging. For the annotation of complete clauses, this paper mainly follows Xavier's idea. First, judge whether the predicted clause head position is multiple clauses, then find out the possible clause candidate set for the starting position of each possible clause, and finally get the most appropriate clause annotation from the scoring function. In order to overcome the problem that the population is easy to fall into local extreme points, the crossover and mutation ideas of genetic algorithm are introduced. The final experimental result is that the particle swarm optimization algorithm converges to the highest recognition accuracy after about 40 iterations. The two literature methods need more than 60 iterations to converge to their highest recognition accuracy. Compared with the two literature methods, particle swarm optimization algorithm has better convergence performance. It is proved that the particle swarm optimization algorithm is feasible for feature recognition and has better convergence performance. In the future work, we will try to effectively introduce more syntactic and semantic features of descriptive clause feature recognition to further improve the result of feature recognition.

References

- [1] Liu Y, Che W, Qin B, et al. Exploring Segment Representations for Neural Semi-Markov Conditional Random Fields. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28:813-824.
- [2] Huebner J, Girardello A, Sliz O, et al. What People Focus on When Reviewing Your App: An Analysis Across App Categories. IEEE Software, 2021, 38(May-June 2021):96-105.
- [3] Zhang Kai, Chen Jianen. Phase Diagram Feature Recognition Method for Multi Defect Parameters of Lattice Sandwich Beam. Computer Simulation, 2022, 39(1):465-470.
- [4] Koak A. Turkish tertiary level EFL learners' recognition of relative clauses. Dilve Dilbilimi Çalışmaları Dergisi, 2020, 16(4):1637-1655.
- [5] Ao Q, Liu N. Multi-feature recognition of English text based on machine learning. Journal of Intelligent and Fuzzy Systems, 2021, 40(2):2145-2156.
- [6] Li C, Raj E D, Namasudra S. Feature Recognition of Abstract Art Painting Multilevel Based Convolutional Ancient Recognition Neural Network Method. Journal of Interconnection Networks, 2021:2141003.
- [7] Abeyrathna K D, Granmo O C, Goodwin M. Extending the Tsetlin Machine with Integer-Weighted Clauses for Increased Interpretability. IEEE Access, 2021, 9:8233-8248.
- [8] Solimine M E. Recognition and enforcement of foreign judgments in American courts and the limits of the law market model. Theoretical Inquiries in Law, 2022, 23(1):97-117.
- [9] Su J C, Choi S W, Kim J H, et al. AI and Text-Mining Applications for Analyzing Contractor's Risk in Invitation to Bid (ITB) and Contracts for Engineering Procurement and Construction (EPC) Projects. Energies, 2021, 14(15):1-28.
- [10] Wang L, Cao X, Li M. English Letter Recognition Based on Tensor Flow Deep Learning. Journal of Physics: Conference Series, 2020, 1627(1):012012 (6pp).

[11] Sl A, Bz A, Sz A, et al. Local discriminant coding based convolutional feature representation for multimodal finger recognition - ScienceDirect. Information Sciences, 2021, 547:1170-1181.

[12] Liu X. Feature Recognition of English Based on Deep Belief Neural Network and Big Data Analysis. Computational Intelligence and Neuroscience, 2021, 2021(6):1-10.

[13] Li P, Jiang S. Analysis of the characteristics of English part of speech based on unsupervised machine learning and image recognition model. Journal of Intelligent and Fuzzy Systems, 2020, 39(99):1-11.

[14] Inunganbi S C, Choudhary P. Recognition of Handwritten Meitei Mayek and English Alphabets Using Combination of Spatial Features// 2020.

[15] Mohammed A J, Al-Khaffaf H. Font Recognition of English Letters Based on Distance Profile Features. Science Journal of University of Zakho, 2020, 8(2):66-71.

[16] Liu D, Tang D, Zhang S, et al. Method for feature analysis and intelligent recognition of infrasound signals of soil landslides. Bulletin of Engineering Geology and the Environment, 2020, 80(3):1-16.

[17] Yue X. Woodworking Tool Wear Condition Monitoring during Milling Based on Power Signals and a Particle Swarm Optimization-Back Propagation Neural Network. Applied Sciences, 2021, 11(19):9026.

[18] Guo L, Zhang X. Adaptive Sliding Mode Active Disturbance Rejection Control of PMLSM Based on Simulated Annealing Particle Swarm Optimization Algorithm// 2021 13th International Symposium on Linear Drives for Industry Applications (LDIA). 2021.

[19] Norwahidayah S, Norania h, N Farahah, et al. Performances of Artificial Neural Network (ANN) and Particle Swarm Optimization (PSO) Using KDD Cup '99 Dataset in Intrusion Detection System (IDS). Journal of Physics: Conference Series, 2021, 1874(1):012061 (8pp).

[20] Dhindsa I S, Gupta R, Agarwal R. Binary Particle Swarm Optimization-Based Feature Selection for Predicting the Class of the Knee Angle from EMG Signals in Lower Limb Movements. Neurophysiology, 2022, 53(2):109-119.

[21] Nabi S, Ahmed M. PSO-RDAL: particle swarm optimization-based resource- and deadline-aware dynamic load balancer for deadline constrained cloud tasks. The Journal of Supercomputing, 2022, 78(4):4624-4654.

[22] An G, Jiang Z, X Cao, et al. Short-term Wind Power Prediction Based On Particle Swarm Optimization-Extreme Learning Machine Model Combined with Adaboost Algorithm. IEEE Access, 2021, 9:94040-94052.

[23] Purushothaman K E, V Nagarajan. Multiobjective optimization based on self, rganizing Particle Swarm Optimization algorithm for massive MIMO 5G wireless network. International Journal of Communication Systems, 2021(5).

[24] Xu J. Online Adaptive PID Control for a Multi-Joint Lower Extremity Exoskeleton System Using Improved Particle Swarm Optimization. Machines, 2021, 10.