2024, 21(2 Special Issue): 0622-0632
https://doi.org/10.21123/bsj.2024.9694
P-ISSN: 2078-8665 - E-ISSN: 2411-7986

Baghdad Science Journal

An Adaptive Harmony Search Part-of-Speech tagger for Square Hmong Corpus

Di-Wen Kang[1], Shao-Qiang Ye[2,4], Sharifah Zarith Rahmah Syed Ahmad[2], Li-Ping Mo[3], Feng Qin[2], Pan Zhou[1]*

[1]School of Communication and Electronic Engineering, Jishou University, Jishou, 416000, China.
[2]Faculty of Computing, Universiti Teknologi Malaysia, Johor, 80310, Malaysia.
[3]College of Computer Science and Engineering, Jishou University, Jishou, Hunan, 416000, China.
[4]College of Information and Engineering, Hunan Applied Technology University, Changde, Hunan, 415000, China.
*Corresponding Author.
ICAC2023: The 4th International Conference on Applied Computing 2023.

Received 27/09/2023, Revised 10/02/2024, Accepted 12/02/2024, Published 25/02/2024

Abstract

Data-driven models perform poorly on part-of-speech tagging problems with the square Hmong language, a low-resource corpus. This paper designs a weight evaluation function to reduce the influence of unknown words. It proposes an improved harmony search algorithm utilizing the roulette and local evaluation strategies for handling the square Hmong part-of-speech tagging problem. The experiment shows that the average accuracy of the proposed model is 6%, 8% more than HMM and BiLSTM-CRF models, respectively. Meanwhile, the average F1 of the proposed model is also 6%, 3% more than HMM and BiLSTM-CRF models, respectively.

Keywords: Harmony Search Algorithm, Low-resource language, Optimization, Part-of-Speech tagging, Unknown words.

Introduction

Through models such as Hidden Markov Model (HMM) and Conditional Random Field (CRF), the goal of part-of-speech (PoS) tagging is to find the most appropriate PoS for all words in the sentence[1-3]. With the development of artificial intelligence technology, the deep learning (DL) models represented by Long Short-Term Memory (LSTM) have been gradually used to solve PoS tagging problems[4]. Due to several reasons, including remoteness in ethnic minority areas, the process of language informatization is slow, resulting in the scarcity of their corpus resources. Some low-resource language PoS tagging solutions are proposed to address the large number of unknown words caused by the scarcity of corpus resources. The approach combines morpheme embeddings, hierarchical Brown clusters, and BiLSTM and selects training data using cross-lingual annotation projection[5]. Magistry et al. used delexicalization and word transformation strategies to calculate the similarity between the target language and a language with richer corpus resources[6]. Li et al. used a mixed Chinese and square Hmong corpus to train the HMM[7]. Because of the significant difference in the data between Chinese and square Hmong languages, the model trained in this way

2024, 21(2 Special Issue): 0622-0632
https://doi.org/10.21123/bsj.2024.9694
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

tends to rely more on Chinese grammar to tag sentences, which may result in incorrect PoS tagging of square Hmong words. The basic idea of the above methods is to use the richer corpus resources language to assist in the PoS tagging of a scarce corpus resources language. However, this approach may lead to semantic mis-understanding[8].

As a population-based optimization algorithm, Harmony Search (HS) mimics the process of musicians searching for the optimal melody to find the optimal value of the objective function[9]. To address the shortcomings of the basic HS algorithm, which tends to get trapped in local optima, has strong randomness in global search, and does not consider all harmonies in the harmony memory during the local search stage, researchers have improved the HS algorithm in two directions: 1) adaptive parameter tuning and 2) integration with other optimization algorithm's search strategies[10]. The improved HS algorithm has been used in various fields, including structural design11, object detection[12], estimating the number of wind turbines[13], and buffer allocation[14].

The PoS tagging problem is an optimization problem[15]. This paper designs a weighted objective function according to the characteristics of the square Hmong corpus. The weight parameters in this function can be adaptively adjusted based on whether the word to be tagged is unknown. Meanwhile, the HS algorithm is improved by utilizing the roulette wheel selection idea and local evaluation strategy to enhance its search capability.

The remainder of this paper contains: The HS algorithm be briefly described in section 2. The PoS tagging objective function and improved HS algorithm are proposed in section 3. Section 4 demonstrates the experiment and analysis. Finally, the entire paper summary be described in section 5.

Harmony Search Algorithm

The HS algorithm seeks optimization by simulating the process of musicians finding optimal harmony. The specific steps are as follows.

Step 1: Initialization. According to the analysis of literature[16], the meanings of each parameter are shown in Table 1.

Table 1. The parameter specific meaning

Parameters	Meaning
HMS	Indicates how many harmonics can be stored in the harmony library.
HMCR	The probability of performing the perturbation operator.
PAR	The probability of fine-tuning.
bw	The step size of fine-tuning.

Step 2: Improvisation. If the condition of random disturbance is satisfied, the *j*-th dimension value of the new solution is generated according to Eq.1.

$$x_j = \begin{cases} x_j + \beta \times bw, & \text{if } 0 < \partial \le 0.5, \\ x_j - \beta \times bw, & \text{otherwise.} \end{cases} \quad 1$$

Where, x_j is the *j*-th dimension value, $\partial, \beta \in (0,1)$.

If the condition of random generation is satisfied, the value of the *j*-th dimension value is generated according to Eq.2.

$$x_j = L + (U - L) \times \beta \quad 2$$

Where, L and U are the minimum and maximum value of the search region, respectively.

Step 3: Update HM. Determine whether to eliminate the worst harmony in HM by comparing the new harmony with the worst harmony.

Step 4: Check termination condition. If meets, the algorithm stops, otherwise, repeat steps 2 to 4.

Convergence analysis

Theorem 1. The optimization process of the HS algorithm can be viewed as an absorbing Markov chain.

2024, 21(2 Special Issue): 0622-0632
https://doi.org/10.21123/bsj.2024.9694
P-ISSN: 2078-8665 - E-ISSN: 2411-7986

Proof: X_{t+1} and X_t represent the population at the t+1 and t iterations, respectively. According to the flow of the HS algorithm, X_{t+1} is formed only under the influence of X_t and the perturbation operator or the restart operator-generated variable $x_t^{'}$, exhibiting the Markov property. Based on the survivor selection mechanism of the population update in HS, have:

$$P\{x_{opt} \notin X_{t+1} | x_{opt} \in X_t\} = 0 \quad 3$$

x_{opt} represents the globally optimal individual. According to the definition of an absorbing Markov chain[17], if a random process X_t satisfies the Markov property and $P\{x_{opt} \notin X_{t+1} | x_{opt} \in X_t\} = 0, t = 1,2,\ldots$ is an absorbing Markov chain, then x_{opt} can be regarded as an absorbing state.

Theorem 2. The HS algorithm may discover the optimal solution in any iteration process.

Proof: Suppose there exists an N-dimensional problem, where each dimension has n possible values, and there is only one optimal solution. The minimum probability of a new individual becoming the optimal solution is:

$$\varepsilon = \sum_{k=0}^{N} C_N^k [(1-HMCR)\cdot \frac{1}{n}]^{N-k}[HMCR \cdot PAR \cdot \frac{1}{n}]^N > 0 \quad 4$$

Eq.3 indicates that the new individual is generated based on old individuals in which each dimension is not optimal, and the probability of the new individual becoming the optimal solution is greater than 0. Therefore, the HS algorithm may discover the optimal solution in any iteration process.

Theorem 3. The HS algorithm always converges as long as the algorithm has a sufficient number of iterations.

Proof: Let $P\{x_t \in X_t\}$ denote the probability of the existence of the optimal solution in X_t, then it suffices to prove $P\{x_t \notin X_t\} = 0$. Have:

$$P\{x_{opt} \notin X_t\} = P\{x_{opt} \notin X_{t+1} | x_{opt} \in X_t\} \cdot P\{x_{opt} \in X_t\} + P\{x_{opt} \notin X_{t+1} | x_{opt} \notin X_t\} \cdot P\{x_{opt} \notin X_t\} \quad 5$$

Combining Eq.3 and Eq.4, have:

$$P\{x_{opt} \notin X_{t+1}\} \le (1-\varepsilon) \cdot P\{x_{opt} \notin X_t\} \quad 6$$

$$0 \le P\{x_{opt} \notin X_{t+1}\} \le (1-\varepsilon)^{t+1} \quad 7$$

$$\lim_{t\to\infty} P\{x_{opt} \notin X_{t+1}\} = 0 \quad 8$$

Proof complete.

□

Improvement ideas for the HS algorithm

According to the convergence analysis, as long as the algorithm has sufficient number of iterations, any improvement algorithm based on the framework of HS algorithm will converge. The search of the HS algorithm is essentially achieved by the operator, which enables the population to move from a local optimal region to a better one, and finally converge to the global optimal region. However, the local optimal point can interfere with the algorithm's ability to find the global optimal solution, which affects the performance of the algorithm in terms of the number of iterations required.

To further analyze why the HS algorithm is prone to falling into local optima, an absorbing Markov chain be established with N states, as shown in Fig.1.

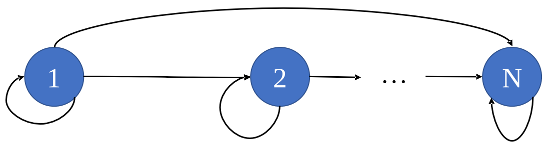

Figure 1. Absorbing Markov chain.

Where, state N is the absorbing state, and these states are arranged from left to right according to their corresponding fitness values. Suppose there is only one individual in the population, and the probability of an individual transitioning from state i to state j is $P_{i\to j}$. Each disturbance operation will only make the state unchanged or move to the right. Based on the steps of the harmony search algorithm, have:

$$P_{i\to k} = 0, k < i \quad 9$$

$$P_{i\to i} = \frac{HMCR \cdot PAR}{2} + \frac{(1-HMCR)\cdot i}{N} + HMCR \cdot (1-PAR) \quad 10$$

2024, 21(2 Special Issue): 0622-0632
https://doi.org/10.21123/bsj.2024.9694
P-ISSN: 2078-8665 - E-ISSN: 2411-7986

$$P_{i\to i+1} = \frac{HMCR \cdot PAR}{2} + \frac{1-HMCR}{N} \quad 11$$

$$P_{i\to j} = \frac{1-HMCR}{N}, j > i+1 \quad 12$$

According to Eq.10, because *HMCR* is less than 1, states closer to the absorbing state are more likely to undergo self-transition, leading to the algorithm falling into local optima. This phenomenon is mainly attributed to the fine-tuning and maintaining operations in the local search part of the harmony search algorithm. The fine-tuning process makes the individual adjustment small and the adjustment direction random, resulting in the quality of the generated new individual not necessarily being better than the previous generation. The maintaining operation increases the possibility of maintaining the state unchanged. Therefore, how to improve the local search part of the HS algorithm to enable the algorithm to efficiently jump from one local optimal region to a better one is a crucial idea for improving the HS algorithm.

Improved harmony search PoS tagger

In this section, according to the characteristics of the square Hmong corpus, the PoS tagging objective function is designed, and an adaptive harmony search tagger (AHST) is proposed. The research framework is shown in Fig.2.

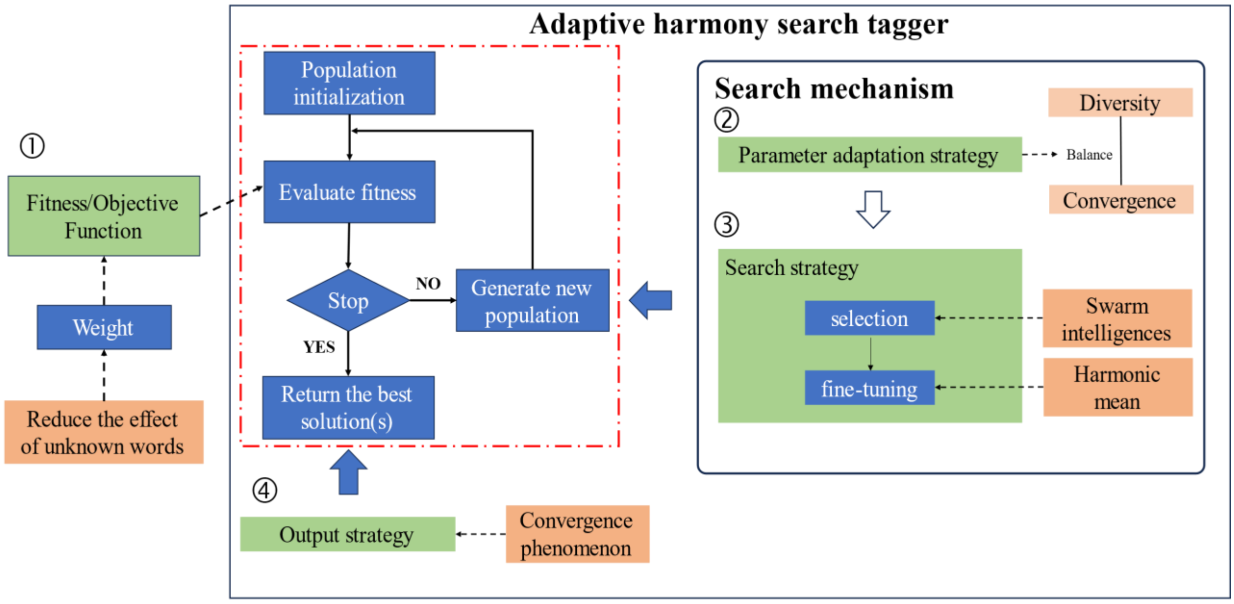

Figure 2. The research framework.

This paper mainly contains the following work: 1) a new objective function is designed to reduce the effect of the unknown words.2) parameter adaptation strategy and search strategy are proposed to improve the search mechanism of HS algorithm. 3) a new output strategy is proposed to consider all candidate solutions.

Objective Function

In PoS tagging problem, HMM and BiLSTM-CRF aim to find the PoS tagging sequence that maximize Eq.13 and Eq.14, respectively.

$$T^* = argmax_{T^j}[\prod_{i=1}^{n} p(w_i|t_i) \times p(t_i|t_{i-1})] \quad 13$$

Where, $p(w_i|t_i)$is the emission probability of the i-th word w_i tagged as t_i, $p(t_i|t_{i-1})$ is the transition probability of the from (i-1)-th tag t_{i-1} to i-th tag t_i.

The objective function Eq.13 is obtained by multiplying and summing the values in the corresponding transfer feature matrix and the emission feature matrix, which makes the value of the objective function sensitive to smaller values. When the value in the feature matrix corresponding

2024, 21(2 Special Issue): 0622-0632
https://doi.org/10.21123/bsj.2024.9694
P-ISSN: 2078-8665 - E-ISSN: 2411-7986

to the correct part of a word in a sentence is small, the correct part of speech is easily eliminated.

$$T^* = \underset{t}{\operatorname{argmax}} score(w,t) \qquad 14$$

$$score(w,t) = \sum_i Emiss(w_i, t_i) + Trans(t_{i-1}, t_i) \qquad 15$$

Where, $Emit(x_i, t_i)$ is the score of the *i*-th word w_i be tagged as t_i. $Trans(t_{i-1}, t_i)$ is the score of the from (*i*-1)-th tag t_{i-1} to i-th tag t_i.

The objective function Eq.15 is obtained by adding the value in the corresponding transfer feature matrix and the value in the emission feature matrix. If the tagged word is an unknown word at this time, the impact of the emission feature and the transfer feature on the objective function is consistent, which may lead to tagging errors.

To reduce the influence of the unknown word, the new objective function is proposed, and shown below.

$$T^* = m_1 * Trans(w_i) + m_2 * Emiss(w_i) \qquad 16$$

Where, $Trans(w_i) = p(t_i|t_{i-1})$ and $Emiss(w_i) = p(w_i|t_i)$. When w_i is the unknown word, $m_1 = 0.99$, $m_2 = 0.01$. This means the $Trans(w_i)$ has a greater impact on the evaluation function than $Emiss(w_i)$, because the part of speech of the unknown word should be determined by the part of speech of the previous word. Otherwise, $m_1 = 0.5$, $m_2 = 0.5$. It means $Trans(w_i)$ and $Emiss(w_i)$ have equal influence. When the previous word is the unknown word, $m_1 = 0.01$, to weak the influence of the PoS of the previous word on the current word tag.

Parameter adaptive adjustment strategy

The parameters of swarm intelligence optimization algorithms affect their performance[18]. In HS algorithm, *HMCR* controls whether to perform the local search or global search, *PAR* controls whether to fine-tune. A larger *HMCR* is beneficial to the local search, and a smaller *HMCR* is beneficial to increasing the diversity of the HM. A larger *PAR* is beneficial to improving the precision, and a smaller PAR makes the algorithm stable. In the early stage of algorithm iteration, global search or fine-tuning should be performed as much as possible. As the algorithm iterates, the frequency of local searches should increase. Therefore, *HMCR* and *PAR* are adjusted according to Eq.17 and Eq.18, respectively.

$$HMCR = 0.3 + 0.69 \times \frac{t}{MAXGEN} \qquad 17$$

$$PAR = 0.99 - 0.49 \times \frac{t}{MAXGEN} \qquad 18$$

Where *t* represents the current number of iterations, *MAXGEN* represents the maximum number of iterations. As the algorithm iterates, *HMCR* will increase from 0.3 to 0.99, and *PAR* will decrease from 0.99 to 0.5.

Search Strategy

Point 1. Harmony Selection Strategy Based on Roulette

A particular part of any harmony in the HM may contain information on the optimal solution. In the basic harmony search algorithm, each iterative evolution does not use all the harmony in the HM, so computing resources are wasted and may lead the algorithm to fall into the local optimum. For the PoS tagging problem, according to the idea of roulette selection, the harmony selecting strategy is further improved, and specific steps are as follows.

Step 1: Calculate the PoS distribution by the corresponding words in the current HM.

Step 2: Math the generated random number with selection probabilities to determine which part to be chosen.

HM[0]	**N**	**D**	**NP**	**W**
HM[1]	**N**	**N**	**NP**	**V**
HM[2]	**V**	**B**	**NP**	**V**

Figure 3. Example harmony selection strategy.

Fig.3 shows the structure of HM, this means there are four words to be tagged in the current sentence. In the current HM, two harmonies are tagging the first word as 'N', and the remaining harmony tags it as 'V'. According to the above steps, for the first word, when the random number is between 0.66 and

2024, 21(2 Special Issue): 0622-0632
https://doi.org/10.21123/bsj.2024.9694
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

1, ‘N’ will be chosen as the initial value of the first dimension of the new harmony.

Point 2. Fine-Tuning Strategy Based on Local Evaluation

The harmonic mean is susceptible to extreme values, and is more affected by minima than maxima. An evaluation function is proposed based on this feature for fine-tuning. The transition probability matrix and emission probability matrix calculated from low-resource corpora are prone to extreme values. Meanwhile, there is no information related to unknown words in the emission probability matrix calculated by statistical methods, so it is usually assumed that the probability of unknown words being tagged for each part of speech is the same. In order to mitigate the impact of the above situation on the model, combining transition probabilities and emission probabilities via weighted harmonic averaging as an evaluation function for fine-tuning, as shown in Eq.19.

$$f(w_i,t)=\frac{\alpha_i+\beta_i}{\frac{\alpha_i}{p(t|T_{i-1})}+\frac{\beta_i}{p(w_i|t)}} \quad 19$$

$$\alpha_i=\begin{cases}0.99, \text{ if } (i\text{-}1)\text{-th word isn't unknow word,}\\ 0.01, \text{ otherwise.}\end{cases} \quad 20$$

$$\beta_i=\begin{cases}0.5, \text{ if } i\text{-th word isn't unknow word,}\\ 0.01, \text{ otherwise.}\end{cases} \quad 21$$

During the fine-tuning process, compare the initial PoS, the most likely PoS for the current word, and the most likely successor PoS to the previous PoS in the context, and select the one with the highest value among the three according to Eq.19 as the new PoS tag for the word being selected.

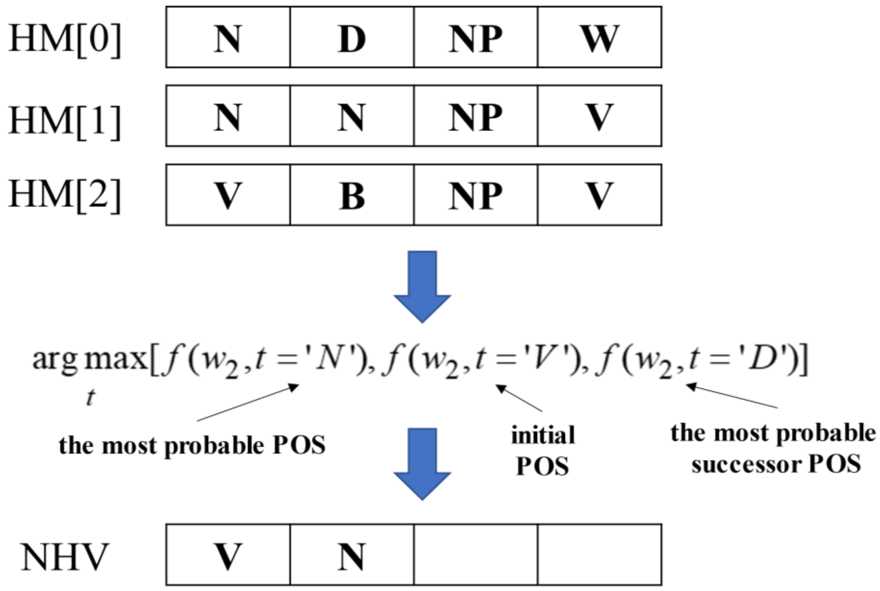

Figure 4. An example of the fine-tuning operation.

As shown in Fig.4, suppose the initial PoS is ‘V’, the most probable POS tag for the current word is ‘N’, and the most probable successor POS tag for the previous tag is ‘D’. If the POS tag with the highest value according to Eq.19 is ‘N’, then the POS tag for the first word in the new harmony NHV will be tagged as ‘N’.

Output strategy

As the algorithm iterates, all solutions converge into a small space, so all the information in the harmony is valuable. When outputting results, the harmony search algorithm only outputs the harmony with the best fitness function in HM as the result, but due to the scarcity of the corpus, the final result cannot be effectively selected only by the fitness function. Therefore, the most frequently occurring values in each dimension in the final HM are combined into a new candidate solution (if the frequency of occurrence is consistent, the value in the optimal harmony is selected), and compared with the

2024, 21(2 Special Issue): 0622-0632
https://doi.org/10.21123/bsj.2024.9694
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

optimal harmony, the optimal fitness function among them is used as the final output result.

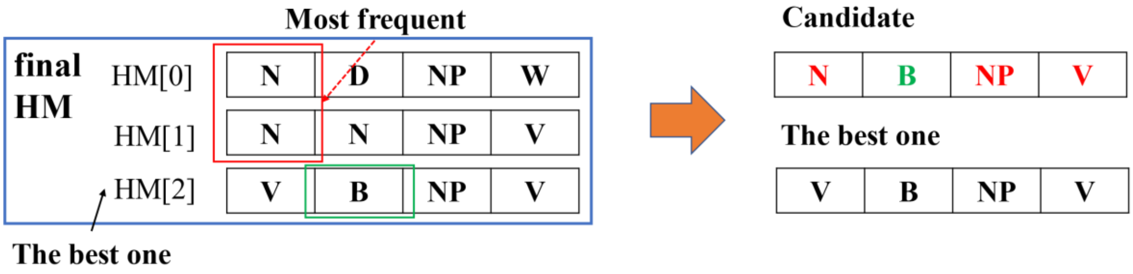

Figure 5. Candidate solution generation process.

As shown in Fig. 5, in final HM, for the first dimension, since N appears the most, the value of the first dimension of the candidate solution is N. For the second dimension, since D, N, and B appear the same number of times, the value of the second dimension of the candidate solution is the value of the second dimension in the optimal harmony, which is B.

Results and Discussion

This section contains the following: Square Hmong corpus description and the experimental results. The experiment conducted environment is Windows 11 operating system and Python 3.7. The hardware specifications used were Intel ® Core™ i5 – 9300H CPU @ 2.40 GHz, 8 GB memory, NVIDIA GeForce GTX 1050 2G.

The Square Hmong Corpus and The Parameter Setting

At the end of the Qing Dynasty, some Hmong words intellectuals created square Miao characters by imitating the structure of Chinese characters to develop Hmong culture. There are two primary sources of the PoS tagging corpus in square Hmong script: the Hmong literature recorded in square Hmong script collected through non-governmental field surveys and collections; the other is script books or papers containing square Hmong script retrieved from the Internet text. The square Hmong PoS tagging corpus used in the experiment has 5942 words. Divide the data set into a training set and a test set in a ratio of 8:2. The average proportion of unknown words in the test set is 44.7%. The evaluation indicators include Accuracy, precision, recall, and F1.

The Experiment results

To determine the parameter value of *HMS* in AHST, 10, 50, and the length of the sentence are chosen as the value of *HMS*. By running ten times independently, the results obtained are as follows.

Table 3. Performance comparison of different *HMS* Values.

HMS		Accuracy	Precision	Recall	F1
10	Best	61.58%	60.82%	61.58%	61.20%
	Worst	59.89%	59.04%	59.89%	59.46%
	Mean	60.64%	59.99%	60.64%	60.31%
50	Best	59.89%	60.18%	59.89%	60.04%
	Worst	58.76%	58.71%	58.76%	58.74%
	Mean	59.57%	59.79%	59.57%	59.68%
the length of the sentence	Best	61.21%	60.43%	61.21%	60.80%
	Worst	60.26%	59.68%	60.26%	60.10%
	Mean	60.64%	60.02%	60.64%	60.30%

According to the experimental results, when *HMS* = 50, the model performance is worse. When *HMS* = 10 or the value of *HMS* equals the length of the sentence, the model performance is the same.

2024, 21(2 Special Issue): 0622-0632
https://doi.org/10.21123/bsj.2024.9694
P-ISSN: 2078-8665 - E-ISSN: 2411-7986

Because *HMS* represents the number of harmonies in the harmony memory, the smaller the *HMS* value, the lower the computational cost.

Ablation experiments were used to verify the effectiveness of each improvement strategy. Strategy 1 represents the parameter adaptive adjustment strategy, strategy 2 represents the search strategy, and Strategy 3 represents the output strategy.

Table 4. Ablation experimental results.

Models		Accuracy	Precision	Recall	F1
Standard HS + Strategy 1	Best	48.96%	64.50%	48.96%	55.70%
	Worst	47.65%	62.46%	47.65%	54.40%
	Mean	48.27%	63.42%	48.27%	54.80%
Standard HS + Strategy 2	Best	49.34%	57.61%	49.34%	53.20%
	Worst	45.20%	52.28%	45.20%	48.50%
	Mean	47.42%	54.85%	47.42%	50.90%
Standard HS + Strategy 3	Best	48.02%	64.64%	48.02%	54.90%
	Worst	45.95%	62.14%	45.95%	52.80%
	Mean	47.33%	63.50%	47.33%	54.20%
Standard HS + Strategy 1, 2	Best	61.02%	60.77%	61.02%	60.90%
	Worst	58.95%	56.80%	58.95%	57.90%
	Mean	59.81%	59.21%	59.81%	59.50%
Standard HS + Strategy 1, 3	Best	47.83%	66.29%	47.83%	55.60%
	Worst	47.63%	66.28%	47.65%	55.40%
	Mean	47.78%	66.29%	47.78%	55.50%
Standard HS + Strategy 2, 3	Best	54.99%	59.23%	54.99%	56.70%
	Worst	51.03%	55.38%	51.03%	53.10%
	Mean	53.35%	57.12%	53.35%	55.20%
AHST	Best	61.58%	60.82%	61.58%	61.20%
	Worst	59.89%	59.04%	59.89%	59.46%
	Mean	60.64%	59.99%	60.64%	60.31%

Strategy 1 controls whether the algorithm performs a local or a global search. Strategy 2 mainly strengthens local search capabilities. When the algorithm does not perform the local search, this strategy does not work. The output strategy considers overall harmony library to output the final result. From the experimental results, it can be seen that each strategy improves the model's ability in different aspects.

To evaluate the performance of the tagging model, accuracy, precision, recall, and F1 are used. Meanwhile, the parameter of the model is set as follows.

Table 5. Parameter setting of each model.

Models	Parameters
HMM	-
BiLSTM-CRF	*batch size*=40, *epoch*=5000, *initial learning rate*=0.01, *factor*=0.5, *patience*=100, *minimum learning*=1e-9
AHST	*MaxGen*=100, *HMS*=10

Where, *MaxGen* is the maximum iterations, *HMS* is the number of harmonies (candidate solution) in HM. When training the HMM, Laplace smoothing is used to solve the zero-probability problem. During the training process of BiLSTM-CRF model, the learning rate decay technique is used. If the evaluation metric of the model does not improve after the patience epoch, the learning rate will be reduced according to the factor.

Table 6. Performance comparison of the taggers on square Hmong corpus.

Models		Accuracy	Precision	Recall	F1
HMM	Best	56.87%	55.28%	56.87%	56.04%
	Worst	52.17%	51.99%	52.17%	52.36%
	Mean	54.20%	54.01%	54.20%	54.10%
BiLSTM-CRF	Best	52.54%	62.67%	52.54%	57.16%
	Worst	52.54%	62.67%	52.54%	57.16%
	Mean	52.54%	62.67%	52.54%	57.16%
AHST	Best	61.58%	60.82%	61.58%	61.20%
	Worst	59.89%	59.04%	59.89%	59.46%
	Mean	60.64%	59.99%	60.64%	60.31%

Table 6 demonstrates the results of comparative experiments. Due to the large number of unknown words in the test set, the smoothing technique used, and the randomness in the swarm intelligence algorithm, the results obtained by HMM and AHST are different in the ten times experiments. During the training stage, the BiLSTM-CRF model randomly selects the feature in the training set as the PoS feature of the unknown word, so the results are relatively stable. Errors caused by unknown words in the tagging process of the Viterbi algorithm will be accumulated, affecting the overall tagging of the sentence. AHST considers both word emission and transfer characteristics and dynamically adjusts relevant weights based on whether a word is unknown, thus affecting the result of fine-tuning. The average accuracy of AHST is 6% more than HMM and 8% more than the BiLSTM-CRF model, and the average F1 of AHST is 6% more than HMM, 3% more than the BiLSTM-CRF model.

Conclusion

As a low-resource corpus for PoS tagging, the PoS tagging of the Square Hmong corpus involves many unknown words. A model based on an improved harmony search algorithm is proposed to reduce the influence of unknown words on PoS tagging. In this model, the harmony selection strategy is enhanced by the idea of roulette. Furthermore, the evaluation function parameters are dynamically adjusted based on whether the word is unknown and fine-tuning PoS based on the local evaluation to reduce the impact of unknown words on the overall tagging process. The experiment reveals that compared to HMM and BiLSTM-CRF models. The proposed model achieves 6%, 8%, 6%, and 3% higher average accuracy and F1 in the square Hmong PoS tagging problem.

In the future, we will further prove the mathematical theory of the weighted objective function and optimize the proposed model to assist the construction of a square Hmong corpus.

Acknowledgment

This work is supported by Scientific Research Project of Education Department of Hunan Province (22C0282, 22B0549, 22B1046, 22C0289); the National Natural Science Foundation of China (62066016); the Xiangxi State Science Program Project-Collaborative control of UAV swarm system based on improved cuckoo algorithm (State Financial Education Index [2022] No.5).

Authors' Declaration

- Conflicts of Interest: None.

- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for re-publication, which is attached to the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee in University of Jishou University.

Authors' Contribution Statement

D.W. K., S.Q. Ye and L.P. M. designed the study. D.W. K. performed the experiments. S. Z. R. S. A., F. Q. and P. Z. analyzed the data. D.W. K. wrote the paper with input from all authors.

References

1. Liu B, Du J, Nie B, et al. Part-of-speech Tagging of Traditional Chinese Medicine Diagnosis Ancient Prose Based on Second-order HMM. Computer Engineering. 2017 Jul;43(7):211-216. Chinese. https://doi.org/10.3969/j.issn.1000-3428.2017.07.035.
2. Warjri S, Pakray P, Lyngdoh S A, et al. Part-of-speech (pos) tagging using conditional random field (crf) model for khasi corpora. Int. J. Speech Technol.. 2021 Jun;24(4):853-864. https://doi.org/10.1007/s10772-021-09860-w
3. Awwalu J, Abdullahi S E Y, Evwiekpaefe A E. Parts of speech tagging: a review of techniques. FJS. 2020 Jun;4(2):712-721. https://doi.org/10.33003/fjs-2020-0402-325
4. AlKhwiter W, Al-Twairesh N. Part-of-speech tagging for Arabic tweets using CRF and Bi-LSTM. Computer Speech & Language. 2021 Aug;65:101138. https://doi.org/10.1016/j.csl.2020.101138
5. Eskander R, Muresan S, Collins M. Unsupervised cross-lingual part-of-speech tagging for truly low-resource scenarios. Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). 2020 Nov;4820-4831. https://doi.org/10.18653/v1/2020.emnlp-main.391
6. Magistry P, Ligozat A L, Rosset S. Exploiting languages proximity for part-of-speech tagging of three French regional languages. LRE. 2019 Apr;53:865-888. https://doi.org/10.1007/s10579-019-09463-7
7. Li H C, Mo L P, Zhou K Q. A Part-Of-Speech Tagging Approach for Chinese-Hmong Mixed Text. IOP Conference Series: Materials Science and Engineering. 2020 Feb;864(1):012064. https://doi.org/10.1088/1757-899X/864/1/012064
8. Heid S, Wever M, Hüllermeier E. Reliable part-of-speech tagging of historical corpora through set-valued prediction. arXiv preprint arXiv:2008.01377. 2020 Aug. https://doi.org/10.48550/arXiv.2008.01377
9. Geem Z W, Kim J H, Loganathan G V. A new heuristic optimization algorithm: harmony search. Simulation. 2001 Feb;76(2):60-68. https://doi.org/10.1177/003754970107600020
10. Qin F, Zain A M, Zhou K Q. Harmony search algorithm and related variants: A systematic review[J]. Swarm Evol. Comput. .2022: 101126. https://doi.org/10.1016/j.swevo.2022.101126
11. Shaqfa M, Orbán Z. Modified parameter-setting-free harmony search (PSFHS) algorithm for optimizing the design of reinforced concrete beams. Struct Multidiscipl Optim .. 2019 Mar;60:999-1019. https://doi.org/10.1007/s00158-019-02252-4
12. Song Y, Pan Q K, Gao L, et al. Improved non-maximum suppression for object detection using harmony search algorithm. Appl. Soft Comput.. 2019 Aug; 81:105478. https://doi.org/10.1016/j.asoc.2019.05.005
13. Christodoulou C A, Vita V, Seritan G C, et al. A harmony search method for the estimation of the optimum number of wind turbines in a wind farm. Energies. 2020 Jun;13(11):2777. https://doi.org/10.3390/en13112777
14. Mistarihi M Z, Okour R A, Magableh G M, et al. Integrating advanced harmony search with fuzzy logic for solving buffer allocation problems. Arab J Sci Eng. 2020 Jan;45:3233-3244. https://doi.org/10.1007/s13369-020-04348-2
15. AL-Jumaili A S. A hybrid method of linguistic and statistical features for Arabic sentiment analysis[J]. Baghdad Sci. J.2020; 17(1 (Suppl.)): 0385-0385. https://doi.org/10.21123/bsj.2020.17.1(Suppl.).0385
16. Kang D W, Mo L P, Wang F L, et al. Adaptive harmony search algorithm utilizing differential evolution and opposition-based learning[J]. Math Biosci Eng. 2021; 18(4): 4226-4246. https://doi.org/10.3934/mbe.2021212
17. WANG Zhongmiao, LIU Jun. Computing the Stationary Distribution of Absorbing Markov Chains with One Eigenvector of Diagonalizable Transition Matrices. Chinese Journal of Applied Probability And Statistics. 2020, 36(2): 123-137. https://doi.org/10.3969/j.issn.1001-4268.2020.02.002

18. Naif O S, Mohammed I J. WOAIP: Wireless Optimization Algorithm for Indoor Placement Based on Binary Particle Swarm Optimization (BPSO)[J]. Baghdad Sci. J.2022;19(3): 0605-0605. https://doi.org/10.21123/bsj.2022.19.3.0605

أداة تمييز جزء من الكلام للبحث عن التناغم التكيفي لمربع لغة همونغ كوربوس

دي وين كانغ[1]، شاو تشيانغ يي[2، 4]، شريفة زاريث رحمة سيد أحمد[2]، لي-بينج مو[3] ، فنغ تشين[2]، وبان تشو[1]

[1]كلية الاتصالات والهندسة الإلكترونية، جامعة جيشو، جيشو، 416000، الصين.
[2]كلية الحاسبات، الجامعة التكنولوجية الماليزية، جوهور، 80310، ماليزيا.
[3]كلية علوم وهندسة الكمبيوتر، جامعة جيشو، جيشو، هونان، 416000، الصين.
[4]كلية المعلومات والهندسة، جامعة هونان للتكنولوجيا التطبيقية، تشانغده، هونان، 415000، الصين.

الخلاصة

يعتبر أداء النماذج المبنية على البيانات ضعيفًا فيما يتعلق بمشكلات وضع علامات على جزء من الكلام لمربع لغة الهمونغ ، وهي مجموعة منخفضة الموارد. تصمم هذه الورقة وظيفة تقييم الوزن لتقليل تأثير الكلمات غير المعروفة. يقترح خوارزمية بحث متناغمة محسنة باستخدام الروليت واستراتيجيات التقييم المحلية للتعامل مع مشكلة وضع علامات على جزء من الكلام الهمونغ. أظهرت التجربة أن متوسط دقة النموذج المقترح هو 6%، 8% أكثر من نماذج HMM وBiLSTM-CRF، على التوالي. وفي الوقت نفسه، فإن متوسط F1 للنموذج المقترح هو أيضًا 6%، أي 3% أكثر من طرازي HMM وBiLSTM-CRF، على التوالي.

الكلمات المفتاحية: خوارزمية البحث المتناغمة، اللغة منخفضة الموارد، التحسين، وضع علامات على جزء من الكلام، الكلمات غير المعروفة.