An Accent Marking Algorithm of English Conversion System Based on Morphological Rules

https://doi.org/10.3991/ijet.v16i01.19717

Yanxia Zhao, Wei Ren
Zhejiang Business College, Hangzhou, China

Zheng Li (✉)
Beijing Union University, Beijing, China
nbl_zheng@buu.edu.cn

Abstract—Facing the English conversion system, the existing accent marking algorithms cannot acquire the morphological rules of English, making the accent marking inaccurate, inefficient, and time-consuming. To solve these problems, this paper puts forward an accent marking algorithm of English conversion system based on morphological rules. Specifically, the English audios in a self-developed English corpus were classified by the speaker classification software based on hidden Markov model, as well as audio classification technology, producing the morphological rules of English. After that, the English accents were marked by the maximum entropy model in the English conversion system. The proposed method was proved accurate and efficient in accent marking through experiments. The research results provide a good reference for marking the accents in English conversion system.

Keywords—Morphological rules, English conversion system, maximum entropy model, audio classification, accent marking

1 Introduction

In parametrically synthesized speech, prosody has an important position, and the accent marking is the basis of high-quality speech synthesis [1]. Thus, it is of great significance to accurately and quickly to mark English accents in the corpus [2]. English accent marking requires a lot of material and manpower, while high-intensity, long-time manual marking is prone to more errors, poorer marking consistency, and higher marking costs. Besides, for the diversified needs of speech synthesis, the speech database needs to adapt to different software and hardware environments. Marking the accent in the English conversion system can reduce the cost of speech synthesis, thereby reducing the cost of building a corpus [3]. For this, related scholars have conducted a lot of related research.

Xu et al. [4] performed a morphological analysis of Uyghur language based on machine translation, in which the stems of the Uyghur language were extracted during the training of the machine translation model and usually regarded as a language

source, while the extracted sentences were taken as the source of the target language to achieve the best translation effect; then, based on the machine translation framework, the morphological analysis of Uyghur language was conducted, but this method increases the time used for accent marking. Zhen and Zhu [5] proposed a word vector-based accent marking algorithm, which expands the marking dictionary through the phonetic approximation calculation function of word vector, and combines this dictionary and the phonetic approximate calculation result to complete English accent marking, but this algorithm cannot obtain English morphological rules, resulting in large errors in the accent marking results and low marking precision. Tang et al. [6] presented an English accent marking algorithm based on reinforcement learning; this algorithm developed a marking dictionary according to the characteristics of English pronunciation, analyzed the accent features, improved the observation information through the long-term memory network, and entered the target word into the reinforcement learning framework, thereby achieving the English accent marking; however, it failed to acquire the morphological rules when marking the accents of many English words, which reduces the marking efficiency of the algorithm.

In view of the problems in the above algorithms, the authors proposed an accent marking algorithm of English conversion system based on morphological rules. First, an English corpus was constructed, and the HHM-based speaker classification software and audio classification technology were applied to classify the English audio in the corpus and obtain English morphological rules. In the English conversion system, the maximum entropy model is used to label the English accent. The results show that the algorithm has high accuracy and marking efficiency. This study provides a reference for the accent marking of the English conversion system [7-9].

2 Morphological Rule Extraction Method

2.1 Corpus construction

Following the development of network media and the Internet, more resources of speech have been available at low or no costs. The text of these voice resources is more accurate, and the recording effect of voice fragments is better [10, 11]. Many open text and voice resources are provided in various network media and Internet websites.

The accent marking algorithm of the English conversion system based on morphological rules used the VOA corpus. VOA is one of the largest news broadcasters in the world, providing rich learning materials for English learners. It's clear in the pronunciations and intonations and rich in resources, easy to download on the Internet, and generalized in the voice of the corpus. VOA corpus is the most primitive corpus, so the voices in the corpus have not been processed. Many audio files exist in the broadcasting process. The public corpus, as the basis for accent marking, contains lots of detailed corpus with manual marking, which requires to extract and process the speech part through the recording to facilitate subsequent English accent marking.

The speech in the VOA corpus isn't marked, but directly obtained from the network with no relevant preprocessing, and the voice and the text are generally aligned [12].

English is an accented language. Its accent is an important super segmental phoneme, which can be divided into accent, secondary accent, and light accent. Stressing a syllable or a word makes it stand out from adjacent syllables. The accent is affected by the four elements of sound length, pitch, intensity, and tone quality. Among them, the pitch in English has the greatest influence on accent of English words, followed by the length, intensity, and quality of the sound. The primary accent of some words in English can be determined according to their morphological structure, so the corresponding morphological rules are formulated based on the characteristics of these words, which can ensure that its marking accuracy is generally higher than the rules obtained through the morphological rules [13]. For words that cannot be identified with the primary accent according to the morphological rules, it is necessary to generate the primary accent marking rules; the secondary accents can also be marked according to the morphological rules [14, 15].

The proposed algorithm classified the English audios using the speaker classification software based on the HMM, as well as the audio classification technology. The identification and segmentation of speech, music and mixed sounds in sound files, and the choice of audio features directly affect the final classification performance. The Gaussian mixture model and the HMM were combined in the proposed algorithm to classify audio. The classification process is shown in Fig. 1.

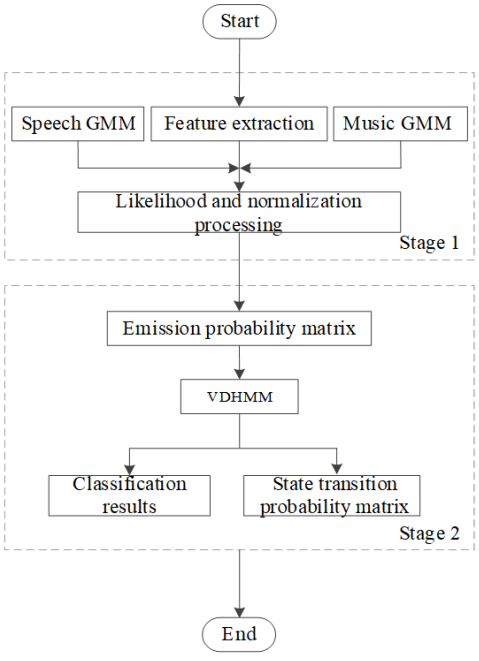

Fig. 1. Flowchart of audio classification

It can be seen from Fig. 1 that the accent marking algorithm is implemented mainly in two stages. In the first stage, analyze and extract the feature parameters of the VOA corpus such as MFCC, short-term energy, spectral parameters, and zero-crossing rate; then use the Gaussian mixture model and the HMM to normalize the likelihood value. In this stage, the utterance, the semantic mixture Gaussian model, and the extracted features work together on the likelihood and normalization. In the second stage, use the state transition matrix of the HMM to calculate the probability value of the observation vector with the obtained likelihood value, and apply the VDHMM to optimally combine the continuous speech frames into segments according to the maximum likelihood criterion, thus completing the audio classification. The final classification result and the state transition probability matrix play an important role in this stage. Among them, the likelihood value and normalization processing, and emission probability matrix are the two most important steps, and both are the basis for connecting the two stages.

2.2 VOA corpus segmentation

After pre-processing, redundant music and noise were removed from the corpus. Then, it's segmented into sentence-based units. First, segment the audio on the basis of HMM unsupervised training. Through iteration, the segmentation result was trained into a more accurate phoneme hidden Markov model in each iteration. Before segmentation, the texts corresponding to the voice should be also processed, because the texts obtained by audio in advance contain many unrecognizable symbols such as @, %, &, *, etc., which are incompatible with the segmentation method and need to be pre-processed.

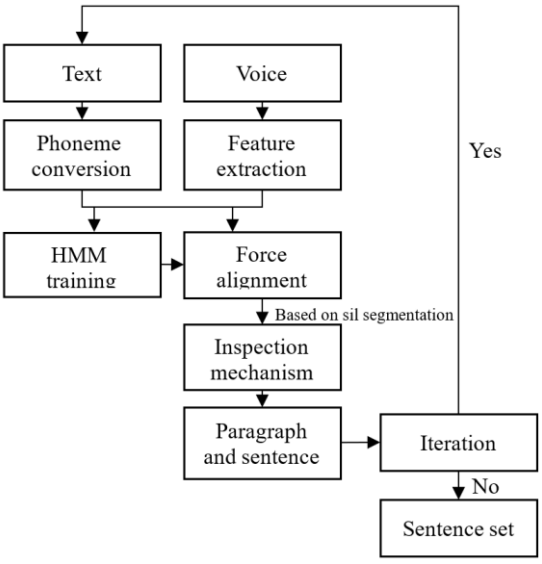

Fig. 2. Flowchart of automatic sentence segmentation

Fig. 2 shows the specific process of automatic sentence segmentation. First, convert the corresponding text in the speech into phoneme sequences, extract the acoustic features from the sound, establish and train the corresponding HMM for each phoneme; then, use forced alignment technology for SIL segmentation. Due to the limitations of the segmentation algorithm, the obtained result was not a collection of sentences and paragraphs, but a collection of sentences. Meanwhile an error detection mechanism was introduced to improve the effect of sentence segmentation, that is, check whether the segmentation point of the sentence is correct, mark the incorrect segmentation point, and repeat the wrong segmentation or sentence for subsequent follow-up elimination of the wrong segmentation part.

2.3 Feature analysis

In the traditional English conversion system, each phoneme is a model, and 40 phonemes are trained as a set of models. Phoneme sets are used to distinguish the accents. Each vowel is divided into two different vowel modes according to the states of accent (Strong accent and light accent). Consonants can be divided into two modes, front and back, to better reflect the syllable structure in words [16]. Acoustic phonological features usually include speech rate, fundamental frequency, pause, duration, phrase boundary, and energy value, etc., which are most related to prosodic information. Prosodic feature information can be obtained based on the above-mentioned acoustic features. The model training result changes with the selected features.

Acoustic features most related to boundary tone and fundamental frequency accent include energy, duration, and fundamental frequency. The pitch change in speech will be affected under the interaction of the three.

Fundamental frequency is an important feature in phonetic notation. English is a stress language, so it's very important for the influence of its prosodic features on the change of fundamental frequency related to English stress. Frequency domain and time domain are the main methods of fundamental frequency detection at present. Time domain features mainly include energy, gene cycle, and zero-crossing rate, while frequency domain features usually include cepstral coefficients, linear prediction forms, and Mel-frequency cepstral coefficients.

The parameters of the fundamental frequency include fundamental frequency value, average value, range, variance and curve.

Let Rn(k) be the fundamental frequency value, which can be calculated by the autocorrelation function:

$$R_n(k) = \sum_{m=0}^{n} S_n(m) S_n(m+k) \tag{1}$$

where, Sn(m) is the parameter vector to be calculated, Sn(m+k) is the number of characteristic functions, and n is the total number of constants in the algorithm.

Let F0-Mea be the mean value of the fundamental frequency, which can be calculated by the number of voiced frames and the sum of the fundamental frequency of voiced frames:

$$F_{0\_Mea} = \frac{\sum_{n=1}^{N} F_{0_n}}{N} \tag{2}$$

where, F_{0_n} is the audio value corresponding to the i-th voiced frame; N is the total number of voiced frames.

The audio range describes the difference in audio, the minimum fundamental frequency value of each word, and the maximum fundamental frequency value of each word.

Let F0-Var be the fundamental frequency variance, then it's calculated as:

$$F_{0\_Var} = \frac{\sum_{n=1}^{N} F_{0_n^2}}{N} - \mu_{F0}^2 \tag{3}$$

where, $F_{0_n^2}$ is the overall coefficient of the fundamental frequency, and μ_{F0}^2 is the basic variance value. In order to accurately realize the gene frequency transformation of each word in the preceding and following words, the current sentence, and the current phrase, it is necessary to obtain the curve contour f0 corresponding to each word based on the above-obtained features [17].

The accent marking algorithm of the English conversion system based on morphological rules adopts the speech analysis software Praat to obtain the duration, energy value, and fundamental frequency value of each word, and then acquire the morphological rules.

3 English Accent Marking Algorithm

The accent marking algorithm of English conversion system based on morphological rules means the use of the maximum entropy model to realize accent marking. The specific process is shown in Fig. 3.

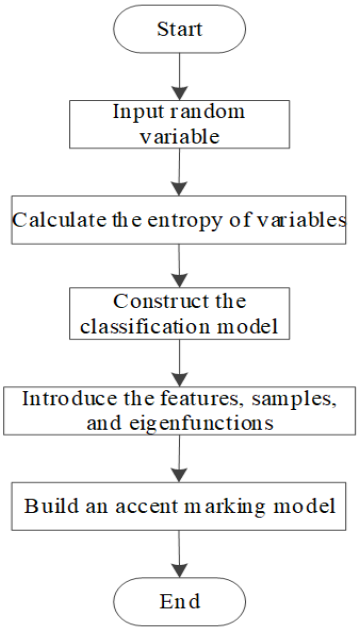

Fig. 3. Flowchart of accent marking

Given that the random variable takes a value in the interval {X=x1, x2, …, xk}, P(X=xi)=Pi represents the probability distribution of the random variable X, and H(X) represents the entropy of X. It's calculated as:

$$H(X) = -\sum_{x \subseteq X} p(x) \log \frac{1}{p(x)} \tag{4}$$

Let H(Y|X) be the conditional entropy of Y if X occurs; P(y|x) be the conditional probability distribution; T be the training data, which is expressed as:

$$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\} \tag{5}$$

This algorithm built a classification model maxH(Y|X) on the basis of the maximum entropy to classify the training data T:

$$\max H(Y \mid X) = \sum_{(x,y)} p(x) \log \frac{1}{p(y \mid x)} \tag{6}$$

When using the maximum entropy model in the modeling process, it's necessary to select features and introduce eigenfunctions, samples and features [18, 19]. P(y|x) indicates features; X is context information; Y is the information that needs to be determined.

Let f(x,y) be the eigenfunction, describing the relationship between output y and input x, and it's expressed as:

$$f(x,y)=\begin{cases}1 & if \;\; y=y_0 \; and \; x=x_0 \\ 0 & otherwise\end{cases} \tag{7}$$

The input value of the eigenfunction f(x, y) to the distribution $\overline{p}(x,y)$ can be calculated as:

$$E_{\overline{p}}(f)=\sum_{x,y}\overline{p}(x,y)f(x,y) \tag{8}$$

The output value of f(x, y) on $\overline{p}(x,y)$ and the model P(Y|X) can be calculated as:

$$E_{p}(f)=\sum_{x,y}\overline{p}(x)p(y\mid x)f(x,y) \tag{9}$$

It's assumed that y is the output result, Y is the set of accent marking, and x is the context of English to be marked. P(y|x) conforming to the context was constructed to achieve the accent marking using the maximum entropy model, and realize the accent marking of the English conversion system according to the marking result.

According to the principle of maximum entropy, P(y|x) needs to satisfy the following conditions:

$$p(y/x)=\frac{1}{Z_{\lambda}(x)}\exp[\sum_{i}\lambda_i f_i(x,y)] \tag{10}$$

$$Z_{\lambda}(x,y)=\sum_{y}\{\exp\sum_{i}\lambda_i f_i(x,y)\} \tag{11}$$

where, Zλ(y/x) represents the normalization factor; λi represents the characteristic parameter.

The accent marking of the English conversion system is not only the basis for the training of English professionals, but also reflects the needs of enterprises for talents. The primary accent is far more important than secondary accent. In the English conversion system, if the conversion of a word is marked correctly, the speech synthesis effect shall be acceptable despite of the deviations in the marking of seconding accent. Primary accent is also simpler than secondary accent. In general, two-syllable words or multi-syllable words have only one primary accent, while secondary accent is more complicated, and the position of secondary accent is often closely related to that of primary accent. The traditional machine learning methods tend to learn primary and secondary accents at the same time, increasing the complexity of learning, and leading to an unsatisfactory learning effect. The proposed algorithm can separate the primary accent and secondary accent for learning, with the aim of improving the precision of primary accent marking as much as possible. To distinguish between primary and

secondary accents, it's assumed that a word has only one primary accent. For a three-syllable word, its primary accent can appear in three positions. If it appears in the first syllable, then it is marked as 0; if in the second syllable, then this syllable is marked as 1.

Let W be the sequence of English words, which is expressed as:

$$W = W_1 W_2 W_3 \cdots W_n \tag{12}$$

Let T be the accent sequence corresponding to the vocabulary sequence, which is expressed as:

$$T = t_1 t_2 t_3 \cdots t_n \tag{13}$$

To study English words, the attributes of the entire word should be extracted, because the primary accent may appear in multiple positions, especially for two-syllable or multi-syllable words. All syllables of the word should be taken into consideration. It's usually believed that the syllable is composed of initial sounds, medium vowel sound, and final sound. Therefore, the initial, vowel sound, and final sounds of all syllables of the word should be extracted as attributes. Each attribute is identified by the first letter of its corresponding English name and the number of the corresponding syllable in the word.

The phonetic pronunciation of each word can be divided into two types: accented and light pronunciation. In the accented pronunciation, the vowels in the stressed syllable are marked as accented, and the remaining vowels are marked as light; in the light pronunciation, all vowels are marked as light. In the syllable, there are consonants before or after the vowel.

On this basis, the statistical analysis of the corpus used in this study showed that the number of word syllables is negatively correlated with its proportion in the corpus, e.g., disyllabic words account for 38.90%, 3-syllable words accounted for 26.55%, and 6-syllable words accounted for 0.6%. Excluding monosyllable words, if all other words are grouped according to the number of syllables, the words with fewer syllables will have fewer attributes, which greatly simplifies the learning process and improves learning efficiency.

The accent marking model was implemented to achieve the accent marking of the English conversion system. For this, it's necessary to calculate the weights of related features, and realize the correspondence between the training model parameter and features, which is conductive to the selection of effective weights and features. The accent marking algorithm of English conversion system based on morphological rules solved the feature vector through GIS algorithm.

4 Experiments and Results

The experiment was conducted on the Linux platform to verify the accent marking algorithm of the English conversion system based on morphological rules.

The precision rates of accent marks was tested for the Algorithm 1 (the morphological analysis of Uyghur language based on machine translation proposed in literature [4]), Algorithm 2 (the word vector-based accent marking algorithm proposed in literature [5]), and Algorithm 3 (the English accent marking algorithm based on reinforcement learning in literature [6]) respectively. It's expressed as:

$$P = \frac{a}{b} \times 100\% \tag{14}$$

where, P is the precision rate of the accent marking; a is the number of correct accent marks of the English conversion system; b is the total number of English words.

The marking precision of Algorithm 1, Algorithm 2, and Algorithm 3 is shown in Fig. 4.

The data analysis in Fig. 4 found that the marking precision obtained in multiple iterations was higher than 80% when it's used to mark the accent of the English conversion system; that of Algorithm 2 and Algorithm 3 fluctuated around 50% and 40% respectively. Comparing the test results of Algorithm 1, Algorithm 2, and Algorithm 3, Algorithm 1 had the highest standard precision rate, because it develops a corpus, and classifies the audios in the corpus using the speaker classification software based on Hidden Markov Model and audio classification technology to acquire the morphological rules, thereby realizing the accent marking of the English conversion system, and improving the marking precision of the algorithm.

In the English conversion system, the efficiency of Chinese and English accent marking is extremely important. The marking time was used as the test index, to perform the accent marking test on Algorithm 1, Algorithm 2 and Algorithm 3. The test results are shown in Fig. 5.

The data analysis in Fig. 5 found that with the increase of English words in the English conversion system, the time used by Algorithm 1, Algorithm 2 and Algorithm 3 to mark the accent continually increased; the time used by Algorithm 1 was lower than that of Algorithm 2 and 3. When the number of words in the English conversion system is as high as 200, the time used for algorithm 2 and algorithm 3 increases significantly because they both haven't obtain English morphological conversion rules, so that they cannot mark a large number of English accents in a short period of time, resulting in longer marking time; whereas, Algorithm 1 uses the maximum entropy model to mark the English accents in the English conversion system based on the English morphological conversion rules, shortening the marking time, and improving the marking efficiency of Algorithm 1.

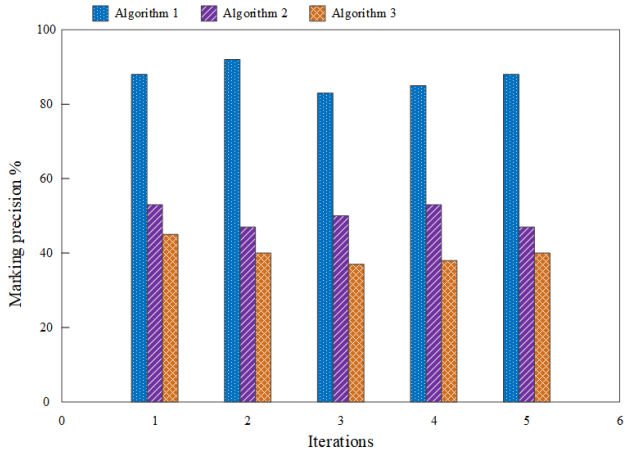

Fig. 4. Marking precision of different algorithms

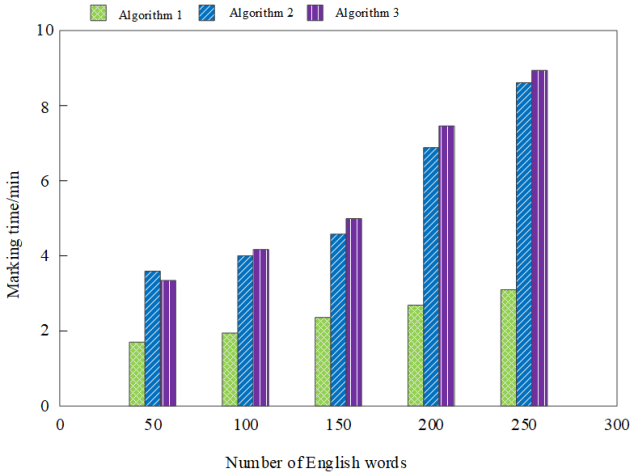

Fig. 5. Accent marking time of different algorithms

5 Conclusion

After ten years of development, the speech synthesis system has been widely used in home network systems and mobile communication equipment fields etc., and people's demands for speech quality have been also continuously increasing under. At this stage, the synthesized speech has a high level of clarity and intelligibility, but its emotional color is not rich enough, the machine taste is heavier, and there is a large gap from human natural language. Therefore, English accent marking of the English conversion system has become a hotspot of research in recent years. The current accent marking algorithm of English conversion system has the problems of low

marking precision and efficiency. To solve these problems, this paper proposes an accent marking algorithm of English conversion system based on morphological rules. An English corpus was developed to obtain morphological rules. On this basis, the maximum entropy model was used to achieve the accent marking of English conversion system, thereby improving the precision and efficiency of marking. This study lays a foundation for the development of the English conversion system and the improvement of the English synthesized speech quality.

6 Acknowledgment

This article is the research result of the project funded by Department of Education of Zhejiang Province (No. Y201942517).

7 References

[1] Zeng, L.P., Chai, P.Q. (2005). English TTS system using variable-length concatenating units. Computer Engineering, 31(3), 180-182. https://doi.org/10.3969/j.issn.1000-3428.2005.03.064

[2] Wang, Y.S., Li, M. (2008). English accent assignment based on morphological rules and machine learning. Journal of Computer Applications, 28(1), 88-91. https://doi.org/10.3724/SP.J.1087.2008.00088

[3] Yang, B., Zhou, L.J., Yu, Z.T., Liu, L.J. (2016). Research on semi-supervised learning-based approach for lao part of speech tagging. Computer Science, 43(9), 103-106. https://doi.org/10.11896/j.issn.1002-137X.2016.9.019

[4] Xu, C., Yang, Y., Jiang, T.H. (2017). Research on machine translation-based Uyghur morphological analysis. Computer Engineering and Applications, 53(14), 138-142. https://doi.org/10.3778/j.issn.1002-8331.1604-0119

[5] Zhen, Y.N., Zhu, J. (2017). A method of Tibetan POS tagging based on distributed representation. Journal of Chinese Information Processing, 31(1), 112-117. http://jcip.cipsc.org.cn/CN/Y2017/V31/I1/112

[6] Tang, S.Q., Sun, Y.R., Li, Z.X., Zhang, C.L. (2020). Part-of-speech tagging of Zhuang based on reinforcement learning. Computer Engineering, 46(4), 309-315. https://doi.org/10.19678/j.issn.1000-3428.0054160

[7] Zhang, T.H. (2020). Recognition and segmentation of English long and short sentences based on machine translation, International Journal of Emerging Technologies in Learning 15(1), 152-162. https://doi.org/10.3991/ijet.v15i101.10182

[8] Zhao, X., Wang, Y.; Liu, Y.L., Xu, Y.N., Meng, Y.N., Guo, L. (2019). Multimedia based teaching platform for English listening in universities, International Journal of Emerging Technologies in Learning, 14(4), 160-168. https://doi.org/10.3991/ijet.v14.i04.9690

[9] Zhang, M. (2020). Virtual situated learning of spoken English based on computer simulation technology, International Journal of Emerging Technologies in Learning, 15(4), 206-217. https://doi.org/10.3991/ijet.v15i04.12939

[10] Wang, X.J., Zhou, L.J., Zhang, J.P., Zhou, F., Guo, J.Y. (2019). Research on the fusion of semi-supervised lao part of speech tagging and word prediction. Journal of Chinese Computer Systems, 40(12), 2500-2505. https://doi.org/10.3969/j.issn.1000-1220.2019.12.005

[11] Liu, S.T., Wang, X. (2020). A research on part-of-speech tagging about Chinese dictionaries in republican period——Taking "Wang Yunwu's Dictionaries" as an example. The Northern Forum, (1), 99-106.
[12] Muhetaer, P., Silamu, W., Maimaitayifu. (2019). Uyghur part of speech tagging method based on hybrid model. Computer Simulation, 36(1), 268-273. https://doi.org/10.3969/j.issn.1006-9348.2019.01.056
[13] Sun, X., Xie, J.S., Wang, R.M. (2017). Neural mechanism of bilingual language transfer. Foreign Language Education, 38(2), 27-32. https://doi.org/10.16362/j.cnki.cn61-1023/h.2017.02.005
[14] Vaxman, A. (2018). The scales-and-parameters approach to morpheme-specific exceptions in accent assignment: Theories, methods and data. The Study of Word Stress and Accent, Cambridge University Press, 387-424. https://doi.org/10.1017/9781316683101.014
[15] Li, X.C. (2016). On the transformation of English and Chinese modal meanings from the perspective of systemic functional linguistics. Social Sciences Hunan, (1), 198-202.
[16] Qian, K., Yang, X.Z., Zhen, M.K., Du, W.Q. (2019). Design of I2S to AES/EBU audio conversion system based on FPGA. Chinese Journal of Electron Devices, 42(4), 984-989. https://doi.org/10.3969/j.issn.1005-9490.2019.04.033
[17] Rosenberg, A., Hirschberg, J. (2009). Detecting pitch accents at the word, syllable and vowel level. NAACL-Short '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, 81-84. https://doi.org/10.3115/1620853.1620878
[18] Si, N.W., Wang, H.J., Li, W., Shan, Y.D., Xie, P.C. (2018). Chinese part-of-speech tagging model using attention-based LSTM. Computer Science, 45(4), 66-70, 82. https://doi.org/10.11896/j.issn.1002-137X.2018.04.009
[19] Liu, H.H., Kong, C., de Bruin, A., Wu, J.J., He, Y.Y. (2020). Interactive influence of self and other language behaviours: Evidence from switching between bilingual production and comprehension. Human Brain Mapping, 41(13), 3720-3736. https://doi.org/10.1002/hbm.25044

8 Authors

Yanxia Zhao is a member of School of International Studies, Zhejiang Business College, Hangzhou, China. She works as a teacher and a scholar specializes in Applied linguistics, language learning and English pronunciation. Email: yxzhaoTina@163.com

Wei Ren is a member of School of E-commerce, Zhejiang Business College, Hangzhou, China. She is a professor dedicated to computer algorithm, information technology and e-commerce research. Email: 1070903990@qq.com.

Zheng Li is a member of College of Applied Science and Technology, Beijing Union University, Beijing, China. She is a vice professor who researches on applied science and technology and education management. Email: nbl_zheng@buu.edu.cn

Article submitted 2020-11-07. Resubmitted 2020-12-12. Final acceptance 2020-12-13. Final version published as submitted by the authors.