IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 January 2023, accepted 6 February 2023, date of publication 9 February 2023, date of current version 15 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3243574

RESEARCH ARTICLE

Multi-Head Self-Attention Gated-Dilated Convolutional Neural Network for Word Sense Disambiguation

CHUN-XIANG ZHANG, YU-LONG ZHANG, AND XUE-YAO GAO
School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China

Corresponding author: Xue-Yao Gao (xueyao_gao@163.com)

This work was supported by the Heilongjiang Provincial Natural Science Foundation of China under Grant LH2022F031.

ABSTRACT Word sense disambiguation (WSD) is to determine correct sense of ambiguous word based on its context. WSD is widely used in text classification, machine translation and information retrieval and so on. WSD accuracy is low because disambiguation features can not cover more language phenomenon and the discriminative ability of WSD model is not high. In order to improve accuracy of simplified Chinese WSD, a WSD model based on multi-head self-attention and gated-dilated convolutional neural network(AGDCNN) is proposed. Ambiguous word is viewed as the center and 4 adjacent lexical units are extracted successively toward the left and right side. Words, parts of speech, and semantic categories in 4 adjacent lexical units are vectorized and the vectorized results are input into gated-dilated convolutional neural network to get discriminative features. Then, multi-head self-attention is adopted to learn the difference and connection among discriminative features fully. Finally, classification weights are output from adaptive average pooling layer. Experiments are conducted on SemEval-2007: Task#5 and SemEval-2021: Task#2. Experimental results show that AGDCNN model has higher accuracy compared with other methods. Our goal is to improve the quality of simplified Chinese WSD as much as possible based on current linguistic resources and machine learning methods. The challenge we face is to extract effective discriminative features and design disambiguation model in high quality. Our novelty lies in that gated-dilated convolution is combined with multi-head self-attention to extract effective discriminative features, and learn their difference and connection from word form, parts of speech, and semantic categories.

INDEX TERMS Word sense disambiguation, multi-head self-attention, convolutional neural network, part of speech, semantic category.

I. INTRODUCTION

In natural language, there are a large number of polysemous words, and WSD aims at determining which sense of ambiguous word should be selected as its correct one in particular context. Word sense disambiguation plays a significant role in machine translation, semantic recognition, information retrieval, and other fields. For example, Chinese word '中医' has 2 semantic categories, including 'practitioner_of_Chinese_medicine' and 'traditional_Chinese_medical_science'. It's necessary to determine meanings of ambiguous word '中医' according to the context.

In simplified Chinese, WSD task is often viewed as text classification. Sentences containing ambiguous word are directly input into WSD model after they are split into words. Jaber input medical text containing abbreviations into BERT model to determine their senses in medicine domain [1]. Kim proposed convolutional neural network for text classification (textCNN), which is the first application of CNN in field of text classification [2]. Kalchbrenner et al. proposed dynamic convolutional neural network(DCNN) and the ideas of wide convolution and K-Max pooling were introduced to achieve good results in various NLP tasks [3]. However, the

The associate editor coordinating the review of this manuscript and approving it for publication was Qichun Zhang.

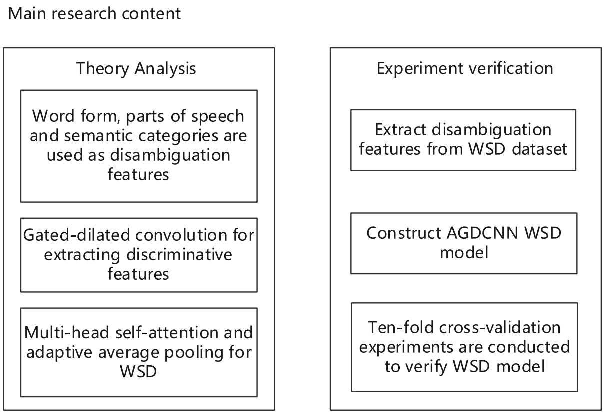

Main research content
Theory Analysis
Word form, parts of speech and semantic categories are used as disambiguation features
Gated-dilated convolution for extracting discriminative features
Multi-head self-attention and adaptive average pooling for WSD
Experiment verification
Extract disambiguation features from WSD dataset
Construct AGDCNN WSD model
Ten-fold cross-validation experiments are conducted to verify WSD model

FIGURE 1. Research scheme of this article.

above methods only consider the singularity of text sentences. Disambiguation features based on words contain the limited information. We should explore features from different perspectives and different dimensions to accomplish WSD task better. In order to achieve the goal, a WSD model based on multi-head self-attention and gated-dilated convolutional neural network (AGDCNN) is proposed.

Therefore, the objective of this paper is to improve accuracy of simplified Chinese WSD compared with current models. The challenge of this paper is how to extract disambiguation features and construct WSD model to accomplish the objective. The novelty of this paper is to use linguistic knowledge to construct disambiguation features, and combine gated-dilated convolution and multi-head self-attention mechanism to extract discriminative features to determine semantic category of ambiguous word. The contributions of this paper are as follows.

(1) This paper takes ambiguous word in Chinese sentence as the center, and selects words, parts of speech and semantic categories in its 4 left and right adjacent lexical units as disambiguation features.

(2) This paper uses gated-dilated convolution to extract discriminative features from words, parts of speech and semantic categories for constructing feature matrix. Gated-dilated convolution not only effectively reduces gradient disappearance but also preserves the nonlinear ability of WSD model.

(3) Multi-head self-attention is used to learn relevant information from different dimensions and representation subspaces, and adaptive average pooling layer is used to determine semantic category of ambiguous word.

The research scheme of this article is shown in Figure 1.

Figure 1 shows research scheme of this paper including theory analysis and experiment verification. Theory analysis includes the definition of disambiguation features, gated-dilated convolution for extracting discriminative features, multi-head self-attention and adaptive averaging pooling for WSD. Experiment verification consists of the extraction of disambiguation features from WSD dataset, AGDCNN WSD model, ten-fold cross-validation experiments.

The remainder of this paper is organized as follows. Section II provides the related work. Section III describes the extraction of WSD features. WSD method based on AGDCNN is given in section IV. Section V gives experimental results and analysis. Section VI provides conclusions and future works.

II. RELATED WORK

A. WORD SENSE DISAMBIGUATION (WSD)

WSD is generally performed in 3 approaches - supervised method, unsupervised one, and semi-supervised one.

In supervised method, labeled data is used to train WSD classifier. Pal et al. extended the baseline strategy in 2019 and gave an improved WSD supervised approach to establish decision trees, support vector machines, artificial neural networks, and naive Bayes models [4]. Chen et al. proposed MetricWSD which transferred knowledge from high-frequency words to infrequent ones by computing distances among senses of a given word through episodic training. As a result, infrequent words and senses enjoyed significant improvement [5]. Barba et al. solved the problem that previous methods did not observe the input sentence and sense definition candidates all at once by redefining WSD as a span extraction problem, which improved the model's performance and generalization power [6]. Du et al.proposed a model of semantic memory for WSD in a meta-learning setting. Semantic memory encapsulated prior experiences seen throughout the lifetime of the model, which aided better generalization in limited data settings [7]. Ranjbar proposed several methods using a pre-trained BERT model, two of which sentences were paraphrased and added as the input to BERT, and one of which WordNet was used to add extra lexical information. Experiments showed that disambiguation accuracy was improved [8]. Sheikh et al. improved the classification model by re-computing logits as a function of both the vanilla independently produced logits and global WordNet graph [9]. Yang et al. focused on enhancing sense representations via incorporating synonyms, example phrases or sentences showing usage of word senses, and sense gloss of hypernyms. Experiments show that incorporating such additional information boosts the performance of WSD [10]. Zheng et al. constructed word formation knowledge to enhance Chinese WSD and proposed FormBERT model, which achieved good results in WSD task [11]. Barba et al. treated WSD as text extraction task and used two Transformer model to improve WSD accuracy [12]. Supervised methods usually have high accuracy. However, training data needs be labeled, and the cost of labeling data is relatively high.

In semi-supervised WSD method, annotated corpus and a large number of non-annotated corpus are used to train WSD classifier. Saqib et al. designed a framework consisting of buzz words in 2018, in which query words had been developed to detect target words based on WordNet. The framework would find the sense of target word using its gloss and examples containing buzz words [13]. Cardellino proposed disjoint semi-supervised learning method, in which unsupervised model was trained on unlabeled data, and the result was used by supervised classifier [14]. Janz explored various expansions to plWordNet as knowledge-bases for WSD and also analyzed the influence of lexical semantic vector models extracted with the help of distributional semantic methods [15]. Sousa et al. adapted semi-supervised algorithms for WSD using word embeddings from word2vec, FastText, and BERT models combined with part-of-speech tags as the input, which achieved better performance [16]. Wei effectively provided attention-driven and long-range dependency for WSD tasks. The proposed PoKED system incorporated position-wise encoding into orthogonal framework and applied knowledge-based attentive neural model to solve WSD problem [17]. Torunoğlu-Selamet et al. proposed semi-supervised approach that used seed sets and context embeddings. He experimented with 9 different contexts based on language models including ELMo, BERT, and RoBERTa, and investigated their impacts on WSD [18]. Hauer et al. combines semi-supervised and unsupervised methods to construct new corpus, which has been experimentally demonstrated to achieve advanced results on standard WSD model [19]. Semi-supervised method provides an effective way to improve WSD model's accuracy using labeled and unlabeled data. But, the cost of a more sophisticated approach is high.

In unsupervised WSD method, untagged corpus is clustered to determine semantic class of ambiguous word. Meng et al. presented context2vec model with POS features to differentiate different meanings represented by one point in vector space [20]. Pesaranghader et al. leveraged bidirectional long short term memory (Bi-LSTM) network that made sense prediction for any ambiguous term. This method also considered a novel technique for automatic collection of training data from PubMed to pre-train the network in an unsupervised manner [21]. Li et al. presented a novel language model based on Bi-LSTM to embed sentential context in continuous space by taking account of word order. They demonstrated that language model can generate high quality context representations in an unsupervised manner [22]. Martin et al. proposed LSA techniques that were deployed as unsupervised learning approach to WSD tasks for sense discovery and distinguishing senses [23]. Hou et al. proposed unsupervised approach for HowNet-based Chinese WSD, which exploited the pre-trained language model [24]. The data of unsupervised method need not be labeled. We can get data easily, but accuracy of WSD is not high.

Maru et al. evaluated previous WSD methods, which carefully analyzed some errors and biases of seven current SOTA models, both qualitatively and quantitatively [25]. Among the above WSD approaches, some of them incorporated WSD task into text classification or text extraction to improve WSD performance, and some of them used different tools at word and sentence vectorization level to improve WSD performance. However, WSD is a tricky problem because it is actually a collection of special and diverse classification problems that are difficult to be clustered in meaningful way. Table 1 shows 3 kinds of WSD methods.

TABLE 1. 3 kinds of WSD methods.

	Labeled corpus	Unlabeled corpus
Supervised WSD	Yes	
Semi-supervised WSD	Yes	Yes
Unsupervised WSD		Yes

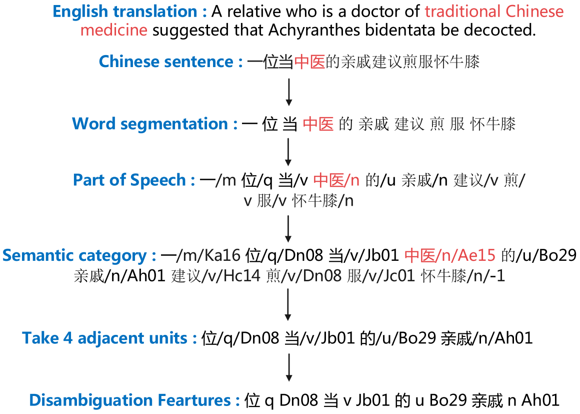
Considering the complexity of semi-supervised method, low accuracy of unsupervised method, and available WSD resources, this paper adopts supervised WSD method.

B. DILATED CONVOLUTION

Dilated convolution is a technique that expanded the kernel by inserting holes between its consecutive elements. It is the same with convolution, but it involves pixel skipping. So, it can cover a larger area of the input. Chen replaced the last few blocks of RESNET with dilated convolution, which made the output size much larger. The amount of computation was not increased, the resolution was not reduced, and a denser feature response was obtained. Details were made better when original image was restored [26]. Wang et al. proposed multi-scale dilated convolution which attempted to handle the obstinate limitation [27]. Receptive field of ordinary convolution is low. Compared with ordinary convolution, dilated convolution increases the interval of scanning features of convolution kernel by changing dilation rate without increasing model parameters. It increases the perceptual field and the output of each convolution contains a larger range of information. Dilated convolution is used to make discriminative features to contain more information in this paper.

C. ATTENTION MECHANISM

Multi-head attention executes attention mechanism several times in parallel. Then, outputs of multiply attentions are concatenated and linearly transformed into the expected dimension. Intuitively, multi-head attention allows for attending to parts of the sequence differently. Zhang et al. gave two WSD classifiers based on bidirectional short-term, and long-term memory networks and self-attention mechanisms in field of biomedicine [28]. Chen et al. presented dynamic convolution, which increased model complexity without increasing the network depth or width. Instead of using single convolution kernel per layer, dynamic convolution aggregated multiple parallel convolution kernels dynamically based upon their attentions [29]. Zhang et al. combined self-attention module and dynamic convolution module by taking a weighted sum of their outputs where weights can be dynamically learned by the model [30]. Zhang et al. proposed a deep-learning acoustic model which used attention mechanism. It used spatiotemporal information and captured emotion-related features more effectively [31]. Feng combines word with part of



English translation : A relative who is a doctor of traditional Chinese medicine suggested that Achyranthes bidentata be decocted.

Chinese sentence : 一位当中医的亲戚建议煎服怀牛膝

Word segmentation : 一 位 当 中医 的 亲戚 建议 煎 服 怀牛膝

Part of Speech : 一/m 位/q 当/v 中医/n 的/u 亲戚/n 建议/v 煎/v 服/v 怀牛膝/n

Semantic category : 一/m/Ka16 位/q/Dn08 当/v/Jb01 中医/n/Ae15 的/u/Bo29 亲戚/n/Ah01 建议/v/Hc14 煎/v/Dn08 服/v/Jc01 怀牛膝/n/-1

Take 4 adjacent units : 位/q/Dn08 当/v/Jb01 的/u/Bo29 亲戚/n/Ah01

Disambiguation Feartures : 位 q Dn08 当 v Jb01 的 u Bo29 亲戚 n Ah01

FIGURE 2. Disambiguation feature extraction.

speech, position and dependency syntax separately to form 3 new combined features and proposed a novel sentiment analysis model based on multi-channel convolutional neural network with multi-head attention mechanism (MCNN-MA) [32]. Self-attention mechanisms can only access information from a single level. Since multi-head attention mechanism no longer uses single attention information, it can learn and represent texts in deeper manner. So, this paper uses multi-head attention mechanism.

In current supervised WSD method, there are two problems. One is that disambiguation feature can not contain more linguistic knowledge and the other is that discriminative ability of WSD model is not high. Therefore, this paper combines words, parts of speech and semantic categories to construct disambiguation features. At the same time, gated-dilated convolution and attention mechanism are fused to improve accuracy of WSD model.

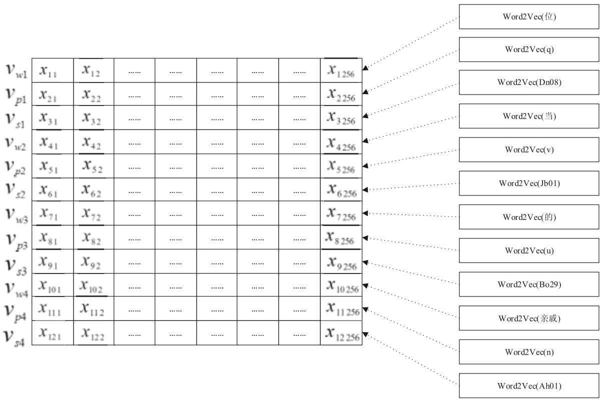
III. WSD FEATURE EXTRACTION

Words closer to ambiguous word have larger influence on its semantics, and words far way from it have less influence on it. Here, sentence containing ambiguous word is segmented. Then, each word is tagged with part of speech and semantic category. Ambiguous word is viewed as the center. Morphologies, parts of speech and semantic categories are extracted as disambiguation features from its 4 left and right adjacent lexical units.

For Chinese sentence '一位当中医的亲戚建议煎服怀牛膝' containing ambiguous word '中医', the process of extracting disambiguation features is shown in Figure 2.

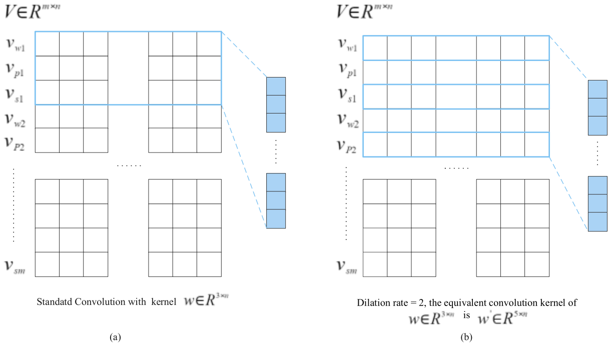
Firstly, Chinese sentence is segmented into words. Secondly, each word is tagged with part of speech. Thirdly, each word is tagged with semantic category. Fourthly, 4 left and right adjacent units around ambiguous word are extracted. Finally, words, parts of speech and semantic categories in these 4 left and right adjacent units are used as disambiguation features.

For the above sentence containing ambiguous word '中医', 12 disambiguation features are extracted. We use Word2Vec tool to vectorize disambiguation features as shown in Figure 3.



v_{w1}
v_{p1}
v_{s1}
v_{w2}
v_{p2}
v_{s2}
v_{w3}
v_{p3}
v_{s3}
v_{w4}
v_{p4}
v_{s4}
x_{11} x_{12} ...... x_{1256}
x_{21} x_{22} ...... x_{2256}
x_{31} x_{32} ...... x_{3256}
x_{41} x_{42} ...... x_{4256}
x_{51} x_{52} ...... x_{5256}
x_{61} x_{62} ...... x_{6256}
x_{71} x_{72} ...... x_{7256}
x_{81} x_{82} ...... x_{8256}
x_{91} x_{92} ...... x_{9256}
x_{101} x_{102} ...... x_{10256}
x_{111} x_{112} ...... x_{11256}
x_{121} x_{122} ...... x_{12256}
Word2Vec(位)
Word2Vec(q)
Word2Vec(Dn08)
Word2Vec(当)
Word2Vec(v)
Word2Vec(Jb01)
Word2Vec(的)
Word2Vec(u)
Word2Vec(Bo29)
Word2Vec(亲戚)
Word2Vec(n)
Word2Vec(Ah01)

FIGURE 3. The generation of word embedding matrix.



$V\in R^{m\times n}$
v_{w1}
v_{p1}
v_{s1}
v_{w2}
v_{P2}
v_{sm}
Standatd Convolution with kernel $w\in R^{3\times n}$
(a)
$V\in R^{m\times n}$
v_{w1}
v_{p1}
v_{s1}
v_{w2}
v_{P2}
v_{sm}
Dilation rate = 2, the equivalent convolution kernel of $w\in R^{3\times n}$ is $w'\in R^{5\times n}$
(b)

FIGURE 4. Structures of convolution and dilation convolution.

IV. WORD SENSE DISAMBIGUATION BASED ON AGDCNN

Convolution kernel continuously scans word embedding matrix V. Dilated convolutions add some spaces (zeros) between convolution kernels and expand the interval of scanning features. It changes dilation rate to increase the receptive field and allows each convolution to output a larger range of information. Convolution with 3×n kernel, dilated convolution with 3×n kernel and 2 dilation rate are shown in Figure 3. In Figure 4(a), convolution kernel is $\mathrm{w}\in\mathrm{R}^{3\times\mathrm{n}}$. In Figure 4(b), dilation convolution kernel is $\mathrm{w}' \in\mathrm{R}^{5\times\mathrm{n}}$.

Matrix $V \in R^{\mathrm{m}\times\mathrm{n}}$ is input to dilated convolution layer. Here, n is the dimension of feature vector. Let $v_i \in R^n$ be the ith row of V, which is feature vector of the ith word. Assuming that dilation rate is d, the height of equivalent convolution kernel is denoted as h calculated in formula (1), which controls the number of words.

$$h = k + (k-1) * (d-1) \tag{1}$$

We use $\mathrm{w}' \in\mathrm{R}^{\mathrm{h}\times\mathrm{n}}$ to perform convolution operation. After convolution operation, feature a_i is extracted as shown in formula (2).

$$a_i = R\left(w' * v_{i:i+h-1}^T + b\right) \tag{2}$$

where, $V_{i:i+h-1} \in\mathrm{R}^{\mathrm{h}\times\mathrm{n}}$ represents feature matrix composed of feature vectors from the ith word to the $(i+h\text{-}1)$th word. Convolution kernel slides on matrix V and step size is 1. With each slide, convolution operation is performed to extract

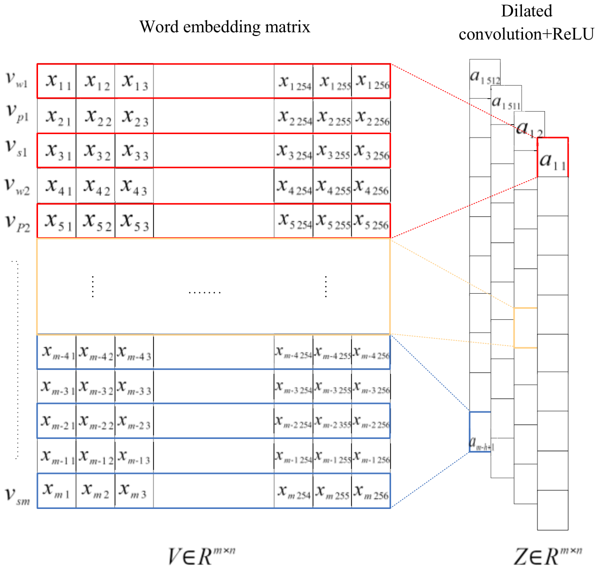

Word embedding matrix
Dilated convolution+ReLU
v_{w1}
v_{p1}
v_{s1}
v_{w2}
v_{P2}
v_{sm}
x_{11} x_{12} x_{13} x_{1254} x_{1255} x_{1256}
x_{21} x_{22} x_{23} x_{2254} x_{2255} x_{2256}
x_{31} x_{32} x_{33} x_{3254} x_{3255} x_{3256}
x_{41} x_{42} x_{43} x_{4254} x_{4255} x_{4256}
x_{51} x_{52} x_{53} x_{5254} x_{5255} x_{5256}
....
x_{m-41} x_{m-42} x_{m-43} x_{m-4254} x_{m-4255} x_{m-4256}
x_{m-31} x_{m-32} x_{m-33} x_{m-3254} x_{m-3255} x_{m-3256}
x_{m-21} x_{m-22} x_{m-23} x_{m-2254} x_{m-2255} x_{m-2256}
x_{m-11} x_{m-12} x_{m-13} x_{m-1254} x_{m-1255} x_{m-1256}
x_{m1} x_{m2} x_{m3} x_{m254} x_{m255} x_{m256}
a_{1512}
a_{1511}
a_{12}
a_{11}
a_{m-h+1}
$V \in R^{m \times n}$
$Z \in R^{m \times n}$

FIGURE 5. The process of convolution operation.

feature until V is traversed. R is nonlinear activation function, and b is the bias. This paper uses ReLU nonlinear activation function as shown in formula (3).

$$R(e) = \max(0, e) \tag{3}$$

where, e represents net activation of convolution layer. When convolution kernel traverses matrix V whose dimension is m, m-h+1 eigenvalues are extracted. Then, eigenvector Z is constructed as shown in formula (4).

$$Z = (a_1, a_2, \ldots\ldots, a_{m-h+1})^T \tag{4}$$

The process of convolution operation on V is shown in Figure 5.

Gated linear unit (GLU) is used to control what information can pass into the following layer. For language modeling task, the gating mechanism allows to select words or features that are important for predicting the next word. Sigmoid function is used to activate Z and σ(Z) is gotten. Hadamard product $\otimes$ between Z and σ(Z) is adopted to compute vector X in hidden layer as shown in formula (5). GLU structure is shown in Figure 6.

$$X = Z \otimes \sigma(Z) \tag{5}$$

We use Hadamard product to multiply corresponding elements of matrix Z and σ(Z) for fusing these features. If we use matrix multiplication, it will lead that wrong results are introduced into vector X. It will bring errors into the subsequent work.

Multi-head attention mechanism is essentially a combination of multiple self-attention mechanisms. In each self-attention mechanism, there is query matrix Q, key matrix K and value matrix V. The output of gated-dilated convolution X becomes initial value of query matrix Q, key matrix K and value matrix V, as shown in formula (6).

$$Q = K = V = X \tag{6}$$

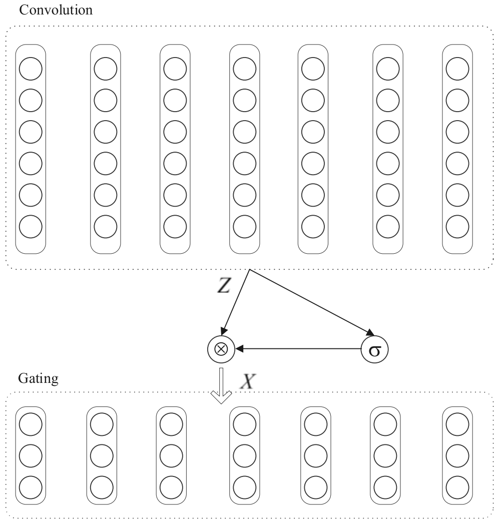


FIGURE 6. GLU structure.

Main idea of self-attention mechanism is scaled dot product attention (SDA). The dot product of Q and K is calculated. Then, the product is divided by $\sqrt{d_k}$. Here, $\mathrm{d_k}$ is the dimension of matrix K and the dot product is not too large. We use softmax function to normalize the result, and multiply it by matrix V to get the score of attention. SDA operation is shown in formula (7).

$$SDA(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

Multi-head attention mechanism uses different parameters w_i^Q, w_i^K, w_i^V to perform linear transformations on matrices Q, K, V in turn, and input linear transformation results into the scaled dot product attention (SDA). The calculation result is represented by $\mathrm{head_i}$, as shown in formula (8).

$$head_i = SDA\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{8}$$

We concatenate the results from $\mathrm{head_1}$ to $\mathrm{head_h}$ to form a matrix, and multiply it by parameter matrix W as shown in formula (9). Here, h is head number.

$$MultiHead(Q, K, V) = Concat(head_i, \ldots, head_h)W \tag{9}$$

The structure of multi-head self-attention is shown in Figure 7.

Adaptive average pooling layer is used instead of fully connected layer. Adaptive average pooling compresses spatial dimension by extracting mean value which is correspondent with the weight of semantic category. Some useless features are compressed to a certain extent. But, fully connected layer has a large number of parameters, and feature maps need be flatten. The comparison between fully connected layer and adaptive average pooling layer is shown in Figure 8.

Ambiguous word c has n semantic classes s_1, s_2, …, s_n. After sentence containing c is input into the proposed WSD classifier, weights $w(s_1|c)$, $w(s_2|c)$, …, $w(s_\mathrm{n}|c)$ are output from adaptive average pooling layer. Semantic class with the

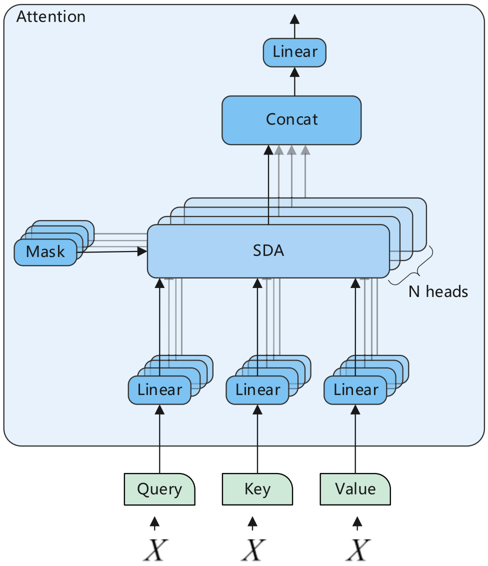


FIGURE 7. The structure of multi-head self-attention.

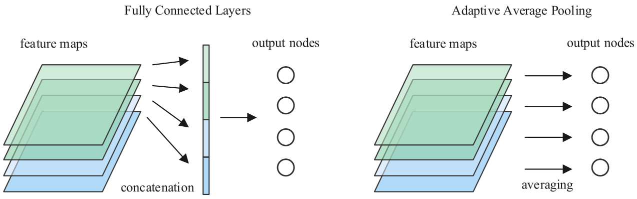


FIGURE 8. Fully connected layer and adaptive average pooling layer.

highest weight is selected as the predicted category, as shown in formula (10)

$$s = \max_{i=1,2,\ldots,n} w(s_i|c) \tag{10}$$

where, s represents the predicted semantic category, and $w(s_i|c)$ is the weight of ambiguous word c under semantic category s_i.

The architecture of AGDCNN-based WSD model consists of dilated convolution layer, GLU layer, multi-head self-attention layer, and adaptive average pooling layer as shown in Figure 9.

V. EXPERIMENTAL RESULTS AND ANALYSES

This paper uses two datasets to verify the performance of AGDCNN. These two datasets are currently the most important datasets for WSD in Simplified Chinese. One dataset is SemEv-al-2007: Task #5. We selected 28 ambiguous words, including 2 categories, 3 categories and 4 categories. The other dataset is SemEval-2021: Task#2. The statistics of two datasets are shown in Table 2.

Experiments are conducted under pytorch deep learning framework. The first group of experiments are conducted to testify the influence of learning rate on WSD. The second group of experiments are performed to testify the influence of attention head number on WSD. The third group of experiments are conducted to investigate the influence of dilation rate on WSD. The fourth group of experiments is to verify the stability of WSD model in which ten-fold cross-validation is used. Precision, recall and F1 are used to evaluate the stability of the proposed network. In the fifth group of experiments, the proposed network is compared with textCNN and MCNN-MA on WSD task.

TABLE 2. Statistics of two data sets.

Data Set	Training data	Test data	Data Size
SemEval-2007: Task #5	2930	515	3445
SemEval-2021: Task #2	1000	1000	2000

TABLE 3. Disambiguation accuracies in the first group of experiments.

Lr	SemEval-2007: Task #5	SemEval-2021: Task #2
0.001	85.08	75.43
0.0001	87.66	78.22
0.00001	59.84	76.21

TABLE 4. Disambiguation accuracies in the second group of experiments.

Heads	SemEval-2007: Task #5	SemEval-2021: Task #2
1	85.36	72.64
2	86.46	74.91
4	87.66	78.22
8	87.04	76.48

Learning rate Lr has a significant impact on AGDCNN. Values of Lr in the first group of experiments are respectively set to 0.001, 0.0001 and 0.00001. Experimental results are shown in Table 3.

The loss curve of ambiguous word '长城' is shown in Figure 10.

From Table 3 and Figure 10, we can see that if learning rate is set too large, the proposed network will wander around optimal value, and the loss will fluctuate or even become the larger. If learning rate is set too small, the proposed network converges slowly. In this case, it falls into local minimum and optimal solution cannot be found. Therefore, learning rate is assigned with appropriate value for ensuring that the loss reaches minimum value faster.

The number of attention heads affects the performance of the proposed network. The second group of experiments aims to investigate the effect of head number on WSD. The number of attention heads is set to 1, 2, 4, and 8. Experimental results are shown in Table 4.

From Table 4, we can see that average accuracy of the proposed network first increases and then decreases with the increase of head number. The proposed network with 4 head attentions achieves the best and its average accuracy reaches 87.66. Different attentions consider the relevance in different

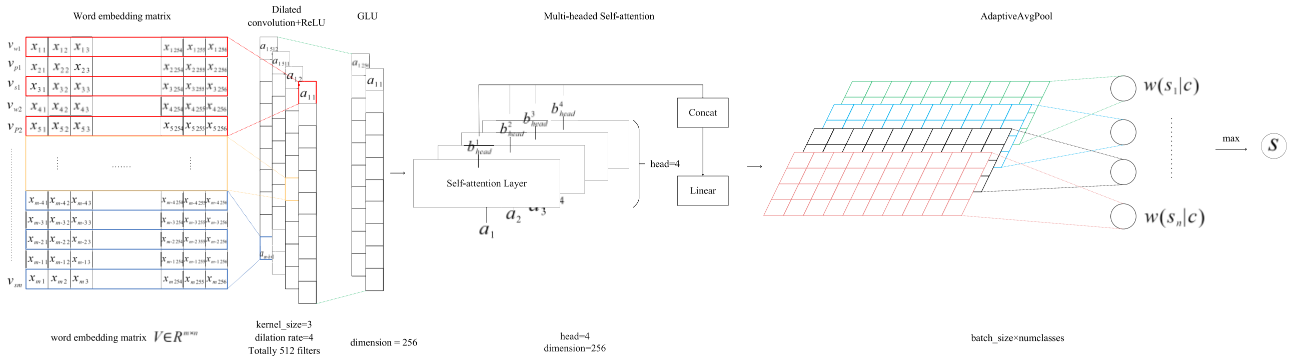


FIGURE 9. The architecture of AGDCNN-based WSD model.

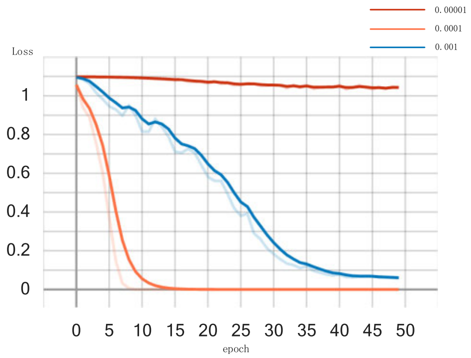


FIGURE 10. Loss curve of '长城'.

TABLE 5. Disambiguation accuracies in the third group of experiments.

Dilation rate	SemEval-2007: Task #5	SemEval-2021: Task #2
1	85.22	75.61
2	86.91	77.00
4	87.66	78.22
6	86.77	75.43
8	84.80	73.86

levels and calculate independently. When head number is smaller, few information is collected and it is difficult to obtain effective features. When head number is larger, information in more levels is considered for obtaining effective features. But, the over-fitting phenomenon occurs easily.

Compared with conventional convolution, dilated convolution extends the scanning range of convolution kernel by setting dilation rate. The convolution kernel is fixed to 3 in the third group of experiments, and dilation rate is set to 1, 2, 4, 6 and 8. Experimental results are shown in Table 5.

From Table 5, we can see that when the height of convolution kernel is set to 3 and dilation rate takes value from set $\{1, 2, 4\}$, average accuracy of the proposed network gradually grows with the increase of dilation rate. This is because that the receptive field is expanded to obtain multi-scale information without losing information and introducing additional parameters. The output of each convolution kernel contains a large range of information. When the height of convolution kernel is set to 3 and dilation rate takes value from set {4, 6, 8}, average accuracy of the proposed network gradually decreases with the increase of dilation rate. This is because that there is local information loss when dilation rate is extremely high. Besides, dilation convolution erases the correlation between the information obtained by long-distance convolution. The reason is that it samples sparsely the input and WSD results are affected. When the height of convolution kernel is set to 3 and dilation rate is set to 4, the proposed network achieves the best and its average accuracy is 87.66.

TABLE 6. Ten-fold cross-validation.

	SemEval-2007: Task #5	SemEval-2021: Task #2
Precision	87.25	77.42
Recall	86.14	76.84
F1	86.69	77.12

So, learning rate is set to 0.0001, head number is set to 4 and dilation rate is set to 4 in the proposed network.

The fourth group of experiments is the validation experiment, in which ten-fold cross-validation is used. Training data are divided into 10 equal parts. 9 parts are used as training set to optimize the proposed network and the rest one part is adopted to testify the optimized network. Precision and recall are used to evaluate the optimized network. Validation experiments are conducted 10 times, in which training set is different in each time. Precision and recall from 10 validation experiments are averaged respectively to verify the stability of WSD model. At the same time, F1 is computed as shown in Table 6.

From Table 6, it can be seen that F1 score reaches 86.69 on SemEval-2007 #Task5 and 77.12 on SemEval-2021: Task #2. This shows that AGDCNN has high stability of word sense disambiguation on these two datasets.

In the fifth group of experiments, the proposed network is compared with WSD model based on textCNN and WSD model based on MCNN-MA. In WSD model based on

TABLE 7. Disambiguation accuracies in the fifth group of experiments.

	SemEval-2007: Task #5	SemEval-2021: Task #2
textCNN	79.61	69.85
AGDCNN	86.20	76.09
MCNN-MA	81.49	73.50

MCNN-MA, convolutional results of each channel are concatenated and input into multi-headed attention layer. Then, the result is input to the fully connected layer to determine category of ambiguous word. In these 3 experiments, words, parts of speech and semantic categories in 4 adjacent left and right units around ambiguous word are used as disambiguation features. Accuracies of these 3 networks on SemEval-2007: Task #5 and SemEval-2021: Task #2 are shown in Table 7.

Table 7 shows that AGDCNN proposed in this paper achieves better than textCNN and MCNN-MA on SemEval-2007: Task #5 and SemEval-2021: Task #2. Accuracy of AGDCNN reaches 86.20% on SemEval-2007 Task #5 dataset, and achieves 76.09% on SemEval-2021: Task #2 dataset. Accuracy of AGDCNN is respectively 4.71% and 2.59% higher than that of MCNN-MA on 2 datasets. This is because that MCNN-MA uses ordinary convolution to extract discriminative features from single layer. AGDCNN adopts gated-dilated convolution to extract discriminative features by increasing the interval of scanning features, which increases the perceptual field and captures multi-scale information. Therefore, accuracy of AGDCNN is higher than that of MCNN-MA. AGDCNN proposed in this paper achieves the best results in the comparison experiments. At the same time, accuracy of AGDCNN is respectively 6.59% and 6.24% higher than that of textCNN on 2 datasets. This proves that AGDCNN is effective in WSD task. This is because that textCNN uses ordinary convolution and AGDCNN utilizes gated-dilated convolution. AGDCNN can extract more effective discriminative features than textCNN. At the same time, AGDCNN adopts multi-head attention to learn the difference between discriminative features. Accuracies of MCNN-MA on SemEval-2007: Task #5 and SemEval-2021: Task #2 are respectively 81.49 and 73.50, which are all higher than those of textCNN. Compared with textCNN, MCNN-MA uses multiple attention mechanism to calculate the similarity between two features for obtaining the connection and difference between them. Therefore, MCNN-MA can extract better discriminative features than textCNN. MCNN-MA achieves better than textCNN on 2 datasets.

Average accuracy of ambiguous words in SemEval-2007: Task #5 and SemEval-2021: Task #2 with the same category number is calculated, as shown in Figure 11.

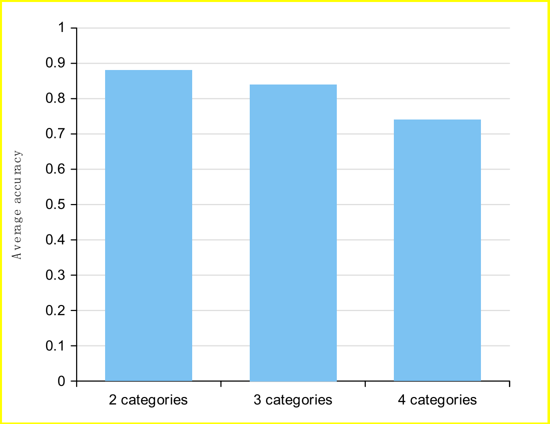


FIGURE 11. Average accuracy under different categories.

From Figure 11, we can find that with category number increasing, average accuracy of ambiguous words with the same category number decreases. This is because that the predicted results have more possibilities when category number increases. It makes error rate of AGDCNN higher. When category number is larger, the distribution of training data of different category is inconsistent. This may lead to data sparsity in training data of some category and average accuracy is decreased. For example, ambiguous word '看' contains 4 semantic categories including 'consider', 'depend on', 'see', 'think'. We find that the probability of correct prediction under category 'think' is less than other 3 categories. This is because that the scale of training data with category 'think' is 50% smaller than other 3 categories. It prevents the proposed network from learning more effective discriminative features under category 'think'. So, WSD accuracy is low under category 'think' and average accuracy of ambiguous word '看' is low.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we propose AGDCNN WSD model for simplified Chinese, which combines gated-dilated convolution and multi-head self-attention mechanism. Disambiguation features are extracted from 4 left and right adjacent words, including words, parts of speech and semantic categories. Gated-dilated convolution is used to extract discriminative features and multi-headed self-attention layer is adopted to learn the difference and connection among discriminative features. Finally, adaptive average pooling layer is used to compute the weight of ambiguous word under each semantic category and category with the largest one is selected. SemEval-2007: Task#5 and SemEval-2021 Task#2 are used to verify the effectiveness of the proposed network. In experiment, head number of multi-head self-attention and dilation rate of gated-dilated convolution are respectively set to 4. Experimental results show that the proposed network achieves higher accuracy than other ones. In the next step, we will consider improving multi-head self-attention mechanism and introducing more linguistic knowledge for improving the performance of WSD model further.

REFERENCES

[1] A. Jaber and P. Martínez, "Disambiguating clinical abbreviations using a one-fits-all classifier based on deep learning techniques," *Methods Inf. Med.*, vol. 61, no. S1, pp. e28–e34, Feb. 2022.

[2] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Oct. 2014, pp. 1746–1751.

[3] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, Baltimore, MD, USA, Jun. 2014, pp. 655–665.

[4] A. R. Pal, D. Saha, N. S. Dash, S. K. Naskar, and A. Pal, "A novel approach to word sense disambiguation in Bengali language using supervised methodology," *Sādhanā*, vol. 44, no. 8, pp. 1–12, Aug. 2019.

[5] H. Chen, M. Xia, and D. Chen, "Non-parametric few-shot learning for word sense disambiguation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 1774–1781.

[6] E. Barba, T. Pasini, and R. Navigli, "ESC: Redesigning WSD with extractive sense comprehension," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 4661–4672.

[7] Y. Du, N. Holla, X. Zhen, C. Snoek, and E. Shutova, "Meta-learning with variational semantic memory for word sense disambiguation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process. (Long Papers)*, vol. 1, 2021, pp. 5254–5268.

[8] N. Ranjbar and H. Zeinali, "Lotus at SemEval-2021 task 2: Combination of BERT and paraphrasing for English word sense disambiguation," in *Proc. 15th Int. Workshop Semantic Eval. (SemEval-)*, 2021, pp. 724–729.

[9] A. El Sheikh, M. Bevilacqua, and R. Navigli, "Integrating personalized PageRank into neural word sense disambiguation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9092–9098.

[10] Y. Song, X. C. Ong, H. T. Ng, and Q. Lin, "Improved word sense disambiguation with enhanced sense representations," in *Proc. Findings Assoc. Comput. Linguistics, (EMNLP)*, 2021, pp. 4311–4320.

[11] H. Zheng, L. Li, D. Dai, D. Chen, T. Liu, X. Sun, and Y. Liu, "Leveraging word-formation knowledge for Chinese word sense disambiguation," in *Proc. Findings Assoc. Comput. Linguistics, (EMNLP)*, 2021, pp. 918–923.

[12] E. Barba, L. Procopio, and R. Navigli, "ExtEnD: Extractive entity disambiguation," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2022, pp. 2478–2488.

[13] S. M. Saqib, F. Masud, A. Hassan, and S. Ahmad, "Semi supervised method for detection of ambiguous word and creation of sense: Using WordNet," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 11, pp. 353–359, 2018.

[14] C. Cardellino and L. A. Alemany, "Exploring the impact of word embeddings for disjoint semisupervised Spanish verb sense disambiguation," *Inteligencia Artif.*, vol. 21, no. 61, pp. 67–81, Mar. 2018.

[15] A. Janz and M. Piasecki, "A weakly supervised word sense disambiguation for Polish using rich lexical resources," *Poznan Stud. Contemp. Linguistics*, vol. 55, no. 2, pp. 339–365, Jun. 2019.

[16] S. Sousa, E. Milios, and L. Berton, "Word sense disambiguation: An evaluation study of semi-supervised approaches with word embeddings," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.

[17] F. Wei and U. T. Nguyen, "PoKED: A semi-supervised system for word sense disambiguation," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2020, pp. 10147–10157.

[18] D. Torunoglu-Selamet, A. Inceoglu, and G. Eryigit, "Preliminary investigation on using semi-supervised contextual word sense disambiguation for data augmentation," in *Proc. 5th Int. Conf. Comput. Sci. Eng. (UBMK)*, Sep. 2020, pp. 337–342.

[19] B. Hauer, "Semi-supervised and unsupervised sense annotation via translations," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process.*, Sep. 2021, pp. 504–513.

[20] Y. G. Meng, "Word sense disambiguation based on context similarity with POS/tagging," *J. Chin. Inf. Process.*, vol. 32, no. 8, pp. 9–18, Aug. 2018.

[21] A. Pesaranghader, S. Matwin, M. Sokolova, and A. Pesaranghader, "DeepBioWSD: Effective deep neural word sense disambiguation of biomedical text data," *J. Amer. Med. Inform. Assoc.*, vol. 26, no. 5, pp. 438–446, May 2019.

[22] Z. Li, F. Yang, and Y. Luo, "Context embedding based on bi-LSTM in semi-supervised biomedical word sense disambiguation," *IEEE Access*, vol. 7, pp. 72928–72935, 2019.

[23] D. I. Martin, M. W. Berry, and J. C. Martin, "Semantic unsupervised learning for word sense disambiguation," in *Supervised and Unsupervised Learning for Data Science*, Sep. 2019, pp. 101–120.

[24] B. Hou, F. Qi, Y. Zang, X. Zhang, Z. Liu, and M. Sun, "Try to substitute: An unsupervised Chinese word sense disambiguation method based on HowNet," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 1752–1757.

[25] M. Maru, S. Conia, M. Bevilacqua, and R. Navigli, "Nibbling at the hard core of word sense disambiguation," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2022, pp. 4724–4737.

[26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[27] Y. Wang, G. Wang, C. Chen, and Z. Pan, "Multi-scale dilated convolution of convolutional neural network for image denoising," *Multimedia Tools Appl.*, vol. 78, no. 14, pp. 19945–19960, Feb. 2019.

[28] C. Zhang, D. Biś, X. Liu, and Z. He, "Biomedical word sense disambiguation with bidirectional long short-term memory and attention-based neural networks," *BMC Bioinf.*, vol. 20, no. S16, p. 502, Dec. 2019.

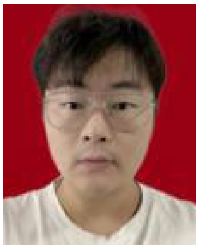
[29] Y. Chen, X. Dai, M. Liu, D. Chen, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11030–11039.

[30] Z. Zhang, S. Wu, G. Chen, and D. Jiang, "Self-attention and dynamic convolution hybrid model for neural machine translation," in *Proc. IEEE Int. Conf. Knowl. Graph (ICKG)*, Aug. 2020, pp. 352–359.

[31] H. Zhang, H. Huang, and H. Han, "Attention-based convolution skip bidirectional long short-term memory network for speech emotion recognition," *IEEE Access*, vol. 9, pp. 5332–5342, 2021.

[32] Y. Feng and Y. Cheng, "Short text sentiment analysis based on multichannel CNN with multi-head attention mechanism," *IEEE Access*, vol. 9, pp. 19854–19863, 2021.

CHUN-XIANG ZHANG received the Ph.D. degree from the MOE-MS Key Laboratory of Natural Language Processing and Speech, School of Computer Science and Technology, Harbin Institute of Technology, in 2007. He is currently a Professor with the School of Computer Science and Technology, Harbin University of Science and Technology. His research interests include natural language processing, machine translation, machine learning, computer graphics and CAD, and 3D model retrieval. He has authored and coauthored more than 60 journals and conference papers in these areas.



YU-LONG ZHANG received the B.S. degree from the School of Computer Science and Technology, Harbin University of Science and Technology, in 2018, where he is currently pursuing the master's degree. His research interests include natural language processing, machine translation, and machine learning.

XUE-YAO GAO received the Ph.D. degree from the School of Computer Science and Technology, Harbin University of Science and Technology, in 2009. She is currently a Ph.D. Supervisor and a Professor with the School of Computer Science and Technology, Harbin University of Science and Technology. Her research interests include computer graphics and CAD, 3D model retrieval, natural language processing, and machine learning. She has authored and coauthored more than 50 journals and conference papers in these areas.