IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 27, 2019, accepted January 13, 2020, date of publication January 30, 2020, date of current version February 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2970436

Hybrid Sense Classification Method for Large-Scale Word Sense Disambiguation

YOONSEOK HEO[1], SANGWOO KANG[2], AND JUNGYUN SEO[1]
[1]Department of Computer Science and Engineering, Sogang University, Seoul 04107, South Korea
[2]Department of Software, Gachon University, Gyeonggi 13120, South Korea

Corresponding author: Sangwoo Kang (swkang@gachon.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Science and ICT under Grant NRF-2019R1C1C1006299.

ABSTRACT Word sense disambiguation (WSD) is a task of determining a reasonable sense of a word in a particular context. Although recent studies have demonstrated some progress in the advancement of neural language models, the scope of research is still such that the senses of several words can only be determined in a few domains. Therefore, it is necessary to move toward developing a highly scalable process that can address a lot of senses occurring in various domains. This paper introduces a new large WSD dataset that is automatically constructed from the Oxford Dictionary, which is widely used as a standard source for the meaning of words. We propose a new WSD model that individually determines the sense of the word in accordance with its part of speech in the context. In addition, we introduce a hybrid sense prediction method that separately classifies the less frequently used senses for achieving a reasonable performance. We have conducted comparative experiments to demonstrate that the proposed method is more reliable compared with the baseline approaches. Also, we investigated the adaptation of the method to a realistic environment with the use of news articles.

INDEX TERMS Computational and artificial intelligence, English vocabulary learning, natural language processing, neural networks, word sense disambiguation.

I. INTRODUCTION

Natural language understanding refers to a series of processes in which a computer reads a text written in human language, analyzes it, and facilitates the reasoning to generate new information [31]. Textual analysis can be divided into a syntactic and a semantic method. A syntactic approach addresses issues such as determining the part of speech in a sentence [18], [21], [22], [42] or identifying the structural relationship of words [6], [9], [17], [40]. These studies have shown considerable promise along with the development of various deep learning models [1], [23], [46]. However, semantic analysis is still challenging at even the most basic level, that of a single word.

In the field of natural language processing, WSD refers to a task that determines a reasonable sense of a word that can have multiple meanings or seems to be ambiguous in a given context. This research still requires more progress, except when the senses of the words can be uniquely determined by the structural features of the sentence. In particular, to commercialize the results of WSD, it will be necessary to address most words and their senses in a wide range of domains; these will range from information retrieval [41], [45], [49] or machine translation [4], [14], [30], [37] to even second language education [7], [50], using various sources such as movies and books.

A dictionary is a very resource that can meet all these needs. The dictionary lists all the senses, their parts of speech, and example sentences for each word that might be used in the real world. Users routinely depend on the dictionary to provide reliable information on the senses of words. However, the senses of the words are very variable, ranging from those that are often encountered in real life to those used only in specific domains such as medicine, natural science, and philosophy.

In this paper, we introduce a new large-scale WSD dataset that is automatically constructed from a widely used and highly reliable source, the Oxford Dictionary. The dataset consists of a set of words and all of the senses described in the dictionary, along with the words' parts of speech uses,

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

and each sense is described in detail and demonstrated with a few example sentences. The key advantage of this data set is that no data construction costs are incurred because the Oxford Dictionary is a publicly accessed resource. In addition, even though the dataset was built automatically, the quality of the data is guaranteed because the dictionary has been created and reviewed by language experts. Lastly, the data used in the existing WSD studies selectively contain only some of the meanings of words, but the current dataset includes all the known meanings of words. Hence, it can be used in various manners irrespective of the domain. Herein we focus on the WSD model for processing this dictionary-based WSD data set.

In trying to understand the sense of a word, the reader makes an appropriate selection based on prior knowledge of the word and the current context. Let's turn this into a machine learning problem. First, we define a ''target word'' as the word for which we want to determine the sense. The semantic decision-making process for a target word involves choosing one sense from among all the senses in the dictionary. The requirement for classification is the contextual information for the word in the sentence.

To address this problem, [16], [39], [48] proposed models for classifying word senses based on traditional machine learning algorithms. These studies define the context information of a target word as the word's part of speech and its surrounding information. This contextual information is then used to develop independent sense classifiers for each word. This approach still shows results competitive with those in other WSD studies but has a scalability problem because a new WSD classifier must be created whenever the number of words in the WSD dataset increases.

On the other hand, the system proposed by [24] assigns the sense of the target word to the one having the most similar context information in the training data by using the K-nearest neighbor algorithm [44]. This study uses the bi-directional LSTM language model to represent the context information of a target word. Following this work, studies by [10], [34] adopted the use of contextual information obtained from the latest neural language models. The performance of such neural language model-based approaches depends on the accuracy with which the contextual information of the target word is represented. The premise for training the neural language models to use different contextual information depending on the sense of words is that the number of example sentences for the sense is sufficient and their distribution is even [13], [38]. In this respect, the datasets used in existing WSD studies satisfy this premise [35]; however, this can be improved in our dataset. Some senses do not contain enough example sentences to train the latest high-capacity neural language models. In addition, while frequently used senses often have a relatively large number of example sentences, there may be very few examples for rarely encountered senses. Because of this balance problem, our dictionary-based dataset makes it difficult for the neural language models to learn the accurate contextual information involving rare senses.

Therefore, in this paper we propose a new WSD model that addresses the greater ambiguity regarding the senses of some words and, at the same time, exhibits greater prediction accuracy for rare senses. The model selects one of the possible senses for the part of speech of the target word in the example sentence. This comes from the format of the Oxford Dictionary, in which all of the senses for each word are recorded and divided according to their parts of speech assignments. This method works well for words with many senses because it reduces the range of sense selection. It also enables neural language models to produce different contextual information for the target word in accordance with its part of speech. In addition, we also propose hybrid sense prediction in which rare senses of a word are classified separately from other, frequently encountered, senses. This has the effect of increasing the prediction accuracy, even if the contextual information for the senses with relatively small numbers of example sentences is somewhat ambiguous.

The remainder of this paper is organized as follows; Section II briefly describes previous studies for WSD, Section III introduces a new large-scale WSD dataset constructed automatically from the Oxford Dictionary, Section IV introduces our proposed WSD model, and Section V discusses our experimental setup and analyzes our results. Finally, we draw conclusions in Section VI.

II. RELATED WORK

Studies on WSD aim to determine the accurate sense of a word in a given context. A line of research like this one can be typically divided into two approaches and may be either supervised or knowledge based. On the one hand, a supervised approach [8], [16], [24], [47] generally makes use of an effective set of features that categorizes senses and creates a classification model that learns the set. To guarantee robust performance, a human-annotated corpus with the labeling of accurate senses of ambiguous words is needed. This prerequisite presents a crucial issue—not only is the construction of such a data set expensive, but it is almost impossible to construct a data set that embraces every sense of every word. Hence, performance is largely dependent on data quality. On the other hand, a knowledge-based approach [3], [5], [19], [28], [32] makes use of an external lexical knowledge base, such as such as[1] WordNet [26] or[2] BabelNet [29], to obtain the sense of an ambiguous word corresponding to its context. The resources generally include sets of words with lexical synonyms grouped into a set, and each set is linked in accordance with the lexical and semantic connections. Thus, when determining the sense of an ambiguous word, the WSD system searches external resources for the sense that suits the context of the word and, thereby disambiguates the word. The major advantage is that these approaches need not depend on human-annotated data in manually curated knowledge resources. However, these

[1] https://wordnet.princeton.edu/
[2] http://babelnet.org

knowledge-based approaches have not yet been shown to outperform the supervised approaches.

Our work belongs to the class of supervised learning approaches. Previous studies are strongly dependent on the ability to properly express the semantic information of words in context [24], [34], [36]. In the past, a word was mapped to its corresponding vector with 1:1 correspondence using a fixed word representation such as Word2Vec [25], GloVe [33], or FastText [2]. For example, in the sentence ''I just dropped by the bank to withdraw money,'' the word ''bank'' itself has multiple meanings such as (1) an organization for financial services and (2) the side of a river or canal. In this case, the fixed word representation always expresses the same embedding of the word ''bank'' regardless of whether its meaning belongs to (1) or (2). In other words, this method always produces vectors that are equally represented, even if they have different meanings in context. Therefore, it is unreasonable to consider the fixed word representation of the ambiguous word in a sentence as contextual information.

Recently, contextualized word embeddings based on a neural language model [10], [24], [34] have been proposed for distinguishably expressing words with different meanings according to contexts, as in the above sentence and meanings (1) and (2). Contextualized word embeddings have been shown to have a significantly positive impact on WSD. [24] have reported important results for a supervised WSD task by using a bi-directional Long Short-Term Memory (BiLSTM) to obtain the contextual information of the target word from the embeddings of the variable-length sentential context around the word. Since then, the latest language models that generate more sophisticated contextualized word embeddings, such as ELMo [34] and BERT [24], have performed successfully in supervised WSD tasks. These models use the k-nearest neighbor algorithm to assign the sense of the target word to the one having the most consistent context information in the training data. Therefore, the success of these approaches depends significantly on the performance of the neural language models. These approaches may be appropriate for the datasets used in the existing studies, where the number of example sentences provided for each sense of the word is sufficient and the distribution among various senses is quite uniform [35]; however, it is problematic to apply a related approach to our dataset, in which some senses do not contain enough example sentences to train the latest high-capacity neural language models. Worse, our dataset has different numbers of example sentences for various senses, depending on their frequency of use in the real environment; hence, the neural language model will encounter difficulty in learning disparate contextual information for rare senses. Therefore, we propose a WSD model that assimilates large amounts of versatile data. In order to address words with many meanings, the model predicts a sense from among those belonging to the target word's part of speech in the example sentence. In addition, the model contains a hybrid sense prediction method and rare senses of a word are classified separately from other, frequently encountered senses.

III. OXFORD DATASET FOR WORD SENSE DISAMBIGUATION

This section introduces a new large data set for WSD that is automatically collected from the publicly accessed[3] Oxford Dictionary. Its primary purpose is to learn a versatile WSD model that can address a wide range of topics, from real-life conversations to machine translation, second language education, medical information retrieval, and more. The key function of a dictionary is to contain all the senses that each word can have in the real world, and they are described in detail and associated with their part of speech tags, lemmas, and example sentences. Also, these content sets are built and reviewed by qualified linguistic experts. Therefore, people generally adopt the dictionary as a standardized resource for determining the meaning of a word.

In this regard, the advantages of utilizing a certified publicly accessed dictionary are as follows; first, no data construction costs are incurred because the dictionary is an open resource. Next, since this dataset contains all the meanings of words in the dictionary, it contains a larger number of senses than do the datasets used in previous studies. Finally, the dataset is sufficiently versatile that it may be used in any domain due to the broad coverage of meanings.

We use a multi-step process to construct our WSD corpus. The Oxford Dictionary contains part of speech, sense ID, sense definition, and example sentences for each word. We first collect words from the dictionary until we have a million example sentences. For even more effective learning, the collection can be preprocessed using the following rules:

- Remove all the example sentences for the word with only one sense ID.
- Remove all the example sentences for the senses not belonging to the list of the predefined part of speech tags.
- For each part of speech tag for a word, the maximum number of senses is limited to 20 and the rest are deleted.
- Remove all the example sentences with more than 50 words or less than 5 words.

In this study, we set up a list of predefined parts of speech tags with nouns, verbs, adjectives, and adverbs for direct comparison with the existing WSD corpus.

Now, we construct the dataset to consist of an example sentence, lemma, part of speech, sense ID, and the position of the target word. We additionally provide a list of parts of speech for the example sentence for use in future work, because a part of speech definition of a word can be still one of the most important features for WSD. To analyze the part of speech tags of the sentence, we adopt[4] Google Cloud NLP.

Examples of our dataset are shown in Figure 1. The first line contains the lemma of the target word, part of speech (referred to in the Oxford Dictionary), sense ID, and word

[3]https://developer.oxforddictionaries.com/
[4]https://cloud.google.com/natural-language/

TABLE 1. Statistics of the Oxford dataset for WSD.

Dataset	# Sentences	# Tokens	# Annotations	# Sense Types	# Lemmas	Ambiguity
Oxford	885,525	18,400,057	885,525	**62,717**	16,444	**14.2**
SemCor	37,176	802,443	226,036	33,362	22,436	6.8
OMSTI	813,798	30,441,386	911,134	3,730	1,149	8.9

```
support Verb    m_en_gbus1015970.008    3
the dome    was supported   by  a   hundred white   columns
DET NOUN    VERB    VERB    ADP DET NUM ADJ NOUN

support Verb    m_en_gbus1015970.015    5
my  main    concern was to  support my  family
PRON    ADJ NOUN    VERB    PRT VERB    PRON    NOUN
```

FIGURE 1. Examples of our Oxford dataset.

position in the sentence from the left. The second line is the sequence of whole words in the sentence, and the third line is the part of speech tags for each word analyzed by Google Cloud NLP.

Table 1 shows the statistics of our Oxford data set compared with others used in previous studies. Two sense-annotated corpora, SemCor [27] and One Million Sense-Tagged Instances (OMSTI) [43], have been widely used in the previous supervised WSD studies. SemCor is the largest human-annotated corpus for WSD and sense annotation comes from the WordNet 1.4 sense inventory. OMSTI is also a large WSD corpus in which every sense is automatically annotated based on the WordNet 3.0 inventory by using an alignment-based WSD approach on a large Chinese-English parallel corpus. Though it can be noisy due to its automatic nature, it has already proven its utility as an additional resource. The major difference of our Oxford data set is that the number of unique senses is the largest of all. In particular, the sense range is almost twice that of SemCor, and significantly larger than that of OMSTI. In addition, we define the ambiguity level as the ratio of the total number of candidate senses to the number of sense annotations, as suggested by [35]. The ambiguity level of our data set is the highest of all, which means that the sense prediction is much more difficult than it is with any other dataset. Additionally, our dataset contains several sentences and is, notably, 25 times larger than that in SemCor.

However, though the corpus has many words with a wide range of sense coverage, it is difficult to use it to train neural language models that can express different contextual information depending on the sense of the word in the context. This stems from the nature of a source (the Dictionary) that does not contain enough example sentences for each sense to train neural language models. Particularly for senses that are less frequently used, the number tends to be extremely low. This differs from other existing datasets, in which the distribution of example sentences for each sense is quite balanced. In the next section, we introduce a novel way of effectively classifying a large number of senses by obtaining distinguishable contextual features from a neural language model while maintaining the advantage of a dictionary dataset.

IV. PROPOSED MODEL FOR WORD SENSE DISAMBIGUATION

In this section, we propose a new WSD model that can effectively address large amounts of versatile WSD data automatically constructed from the Oxford Dictionary. The Dictionary defines all the senses that each word can have in the real world, so each word has a higher level of sense ambiguity than do other datasets. Moreover, the dictionary has every sense of the word classified according to its parts of speech. Given this characteristic, we propose a model that individually predicts the sense of the word according to its part of speech in a context, as can be shown in Figure 2. It facilitates the processing of words with high ambiguity. Additionally, we introduce a hybrid sense prediction that separately predicts less frequently used senses for better prediction accuracy. For each sense of a word, the number of example sentences varies according to its frequency of use in the real environment. As a result, it is difficult for a neural language model to produce distinct contextual information for those senses. Given an example sentence, the model predicts the sense of the target word using the following three steps: (1) it first produces a contextual embedding that represents the contextual information for the target word using a bi-directional LSTM, (2) it uses a switching mechanism to provide a context feature dependent on the part of speech of the target word in the example sentence, and (3) it executes the hybrid sense prediction by calculating the likelihood of rare senses separately when the preceding prediction indicates that the sense belongs to <Rare>, which is a symbol for "a rare sense." In the following section we describe the LSTM, which is the basic element of the neural language model in this paper. Section IV.B explains the method of producing contextual information, Section IV.C introduces the method of generating a context based on the part of speech assignment, and Section IV.D describes the hybrid sense prediction method in detail.

A. LONG SHORT TERM MEMORY

A recurrent neural network (RNN) model is an effective deep neural network that predicts a time sequence. However, the longer the sequence, the higher is the chance of encountering vanishing and exploding gradients; this makes accurate prediction difficult. To solve this issue, a LSTM

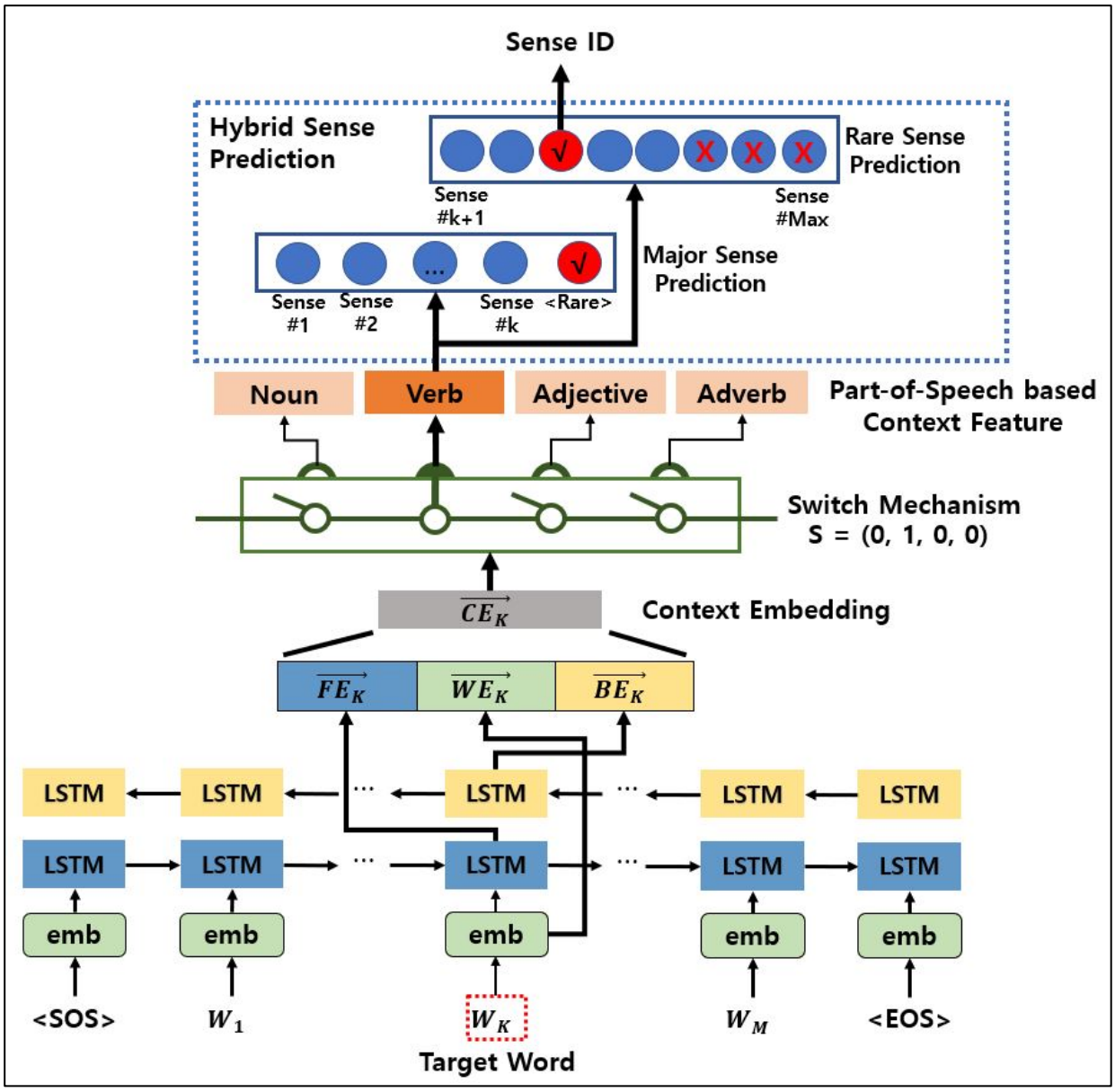

FIGURE 2. Proposed model.

model is presented as a variant structure of RNNs. The LSTM consists of a memory cell to save data, an input gate to receive data, an output gate to export data, and a forget gate to delete data [12], [15]. The structure of LSTM is shown in formulas (1) to (6).

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1}), \quad (1)$$
$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1}), \quad (2)$$
$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1}), \quad (3)$$
$$c'_t = \tanh(W_{cx}x_t + W_{ch}h_{t-1}), \quad (4)$$
$$c_t = x_t \odot c_{t-1} + i_t \odot c'_t, \quad (5)$$
$$h_t = o_t \odot c_t \quad (6)$$

where i_t, o_t, f_t, and c_t denote the input gate, the output gate, the forget gate, and the memory cell, respectively. The memory cell is updated by each gate. Finally, each LSTM cell exports h_t, i.e., the hidden layer, via the output gate.

B. CONTEXT EMBEDDING

In order to obtain the contextual information for the target word in the example sentence, we introduce a context embedding ($\overrightarrow{CE_k}$) that combines the outputs of the Bi-directional LSTM with pre-trained word embedding for the target word.

A target word (W_K) is first represented as a word embedding ($\overrightarrow{WE_k}$) by a pre-trained word embedding model. We particularly exploit the GloVe that can be obtained from unsupervised learning via co-occurrence statistics comparing words in a corpus [33]. The GloVe model in this paper was pre-trained with Wikipedia 2014 and Gigaword 5.

Then, contextual information for the target word can be obtained from the output of BiLSTM. Given a sequence of M words ($W_1, W_2, \ldots, W_M$), a forward LSTM aggregates all the contextual information from the start (<SOS>) to the Kth step which corresponds to the target word and generates the output vector ($\overrightarrow{FE_k}$) from the last output. Similarly, a backward LSTM executes the same process except it works from the end (<EOS>) to the target, and it generates the output vector ($\overrightarrow{BE_k}$). Next, all three vectors are concatenated into one ($\overrightarrow{O_k}$) to generate the final context embedding vector.

$$\overrightarrow{O_k} = [\overrightarrow{FE_k}; \overrightarrow{WE_k}; \overrightarrow{BE_k}] \tag{7}$$

Finally, a feed forward network is used to produce the context embedding ($\overrightarrow{CE_k}$) from the concatenated vector ($\overrightarrow{O_k}$):

$$\overrightarrow{CE_k} = FFN(\overrightarrow{O_k}) \tag{8}$$

$$FFN(x) = ReLU(L_1(x)) \tag{9}$$

where FFN stands for Feed Forward Neural Network, ReLU is the Rectified Linear Unit [11], and $L_1(x) = W_1x + b_1$. We set as 300 the dimensionality of the word embedding, the hidden units of the BiLSTM, and the hidden size of FFN. Consequently, the final context embedding ($\overrightarrow{CE_k}$) of the target word can be interpreted as a contextual meaning representation for the entire input sentence.

C. PART OF SPEECH BASED CONTEXT FEATURE

In order to address effectively words with high sense ambiguity, we add an additional layer to obtain a context feature dependent on the part of speech of the target word. It enables the user to predict the sense of the word according to its part of speech in a context. This layer contains independent feed forward networks (FFN_i) along with the parts of speech designations for the word. We employ a switch to activate a certain network determined by the part of speech in the input sentence. We adopt the part of speech of the target word in the sentence from the Dictionary. If the part of speech of the target word is a verb, as shown in Figure 2, the feed forward network connected to the second switch is then activated. Now, each network in this layer uses as input context embedding created in the previous step to generate a part of speech context feature ($\overrightarrow{F}$) to predict the sense:

$$\overrightarrow{F_i} = FFN_i(\overrightarrow{CE_k}), \quad i \in \{Noun, Verb, Adjective, Adverb\} \tag{10}$$

$$FFN_i(x) = ReLU(L_{2i}(x)) \tag{11}$$

where FFN stands for Feed Forward Neural Network, ReLU is the Rectified Linear Unit, and $L_{2i}(x) = W_{2i}x + b_{2i}$. We set the hidden size for each feed forward network as the maximum number of meanings a word has for a certain part of speech in the training set.

D. HYBRID SENSE PREDICTION

To alleviate the problem that the neural language model is unable to generate distinct contextual information for rare senses which contain a relatively small number of example sentences, we introduce a hybrid sense prediction method that predicts the rare senses separately. This method originally comes from the field of neural machine translation. A hybrid neural machine translation model proposed by [20] has been shown to improve the translation for low frequency or unknown words in the training data, which are usually symbolized as <UNK>. In the field of machine translation, there is an ongoing debate over how to determine a certain word identified as an <UNK> word in a decoding step. This stems from the difficulty in training a neural network on the syntactic and semantic alignment for the less frequently seen words or those that do not exist in the training data.

Motivated by this approach, our hybrid sense prediction proceeds in two phases. The first phrase is major sense prediction, which determines whether the sense of the target word is rare or not using the part of speech context feature ($\vec{F}$). More specifically, this step predicts whether the word's sense belongs to one of the top K senses with high frequency, or to the group of rare senses symbolized as <RARE> in Fig. 2. We set a feed forward neural network (FFN_{Major}), and the context feature ($\overrightarrow{F}$) is given as an input. The likelihood of sense categories including <RARE> is calculated as follows:

$$\overrightarrow{MS} = FFN_{Major}(\overrightarrow{F}) \tag{12}$$

$$\mathrm{FFN}_{Major}(x) = Softmax(ReLU(\mathrm{L}_{3i}(x))) \tag{13}$$

where ReLU is the Rectified Linear Unit and $L_{3i}(x) = W_{3i}x + b_{3i}$. We set the hidden size of the feed forward network as K+1, where K is the number of senses that are frequently used. We set its value as 4, which is determined experimentally. If the maximum likelihood belongs to one of the senses other than <RARE>, then that sense is assigned as the sense of the target word. Otherwise, it belongs to <RARE> and the model then goes on to the second phase.

The second phase involves rare sense prediction. If the maximum likelihood of senses in the first phase lies in the <RARE> category, sense prediction is performed in the following additional neural network. We add a feed forward network (FFN_{Rare}) that is totally independent of the one used in the first step solely for identifying rare senses. The likelihood of rare sense categories is calculated as follows:

$$\overrightarrow{RS} = FFN_{Rare}(\overrightarrow{F}) \tag{14}$$

$$FFN_{Rare}(x) = Softmax(ReLU(L_{4i}(x))) \tag{15}$$

where $L_{4i}(x) = W_{4i}x + b_{4i}$. We set the hidden size of the feed forward network as the number of rare senses. The sense would then be assigned to the one with the maximum likelihood. In order to train the overall network, we set our objective function as follows:

$$J = J_{Major} + \alpha J_{Rare} \tag{16}$$

where J_{Major} and J_{Rare} represent the loss of the major sense prediction and the rare sense prediction, respectively. In this study, each loss function is the cross-entropy, and alpha is set to 1.

V. EXPERIMENT

This section consists of two parts. Section V.A describes the dataset and training details used in the experimental setup. We show the performance of our proposed model and investigate the results in Section V.B.

A. EXPERIMENTAL SETUP

We developed our new dataset for WSD from the publicly available Oxford Dictionary, as mentioned in Section 3. The dataset consists of 765,145 sentences, each of which contains an ambiguous word annotated with its sense and part of speech. The training, validation, and test sets consist of 624,281, 70,864, and 70,000 sentences, respectively.

We also built another test set, called News Test, to verify the performance of the proposed model in a more realistic environment. We collected 3,072 news articles from The Wall Street Journal, selected random vocabulary, and personally tagged relevant sense IDs.

The hyperparameters used in the proposed model are shown in Table 2. We trained our model until the accuracy of the validation set does not improve for 5-epochs. In addition, we adopted the baseline as the context2vec and its hyper parameter is set to match that in the newspaper.

TABLE 2. Hyperparameters.

Hyperparameter	Value
Word embedding Size	300
LSTM hidden size	300
Feature layer size	300
Maximum sense categories per pos	15
Max sequence length	50
Dropout	0.3
Batch size	128
Learning rate	0.0008
Loss Function	Cross-Entropy

B. EXPERIMENTAL RESULT

In this section, we have investigated that our proposed model can effectively address many words with a wide range of sense coverage compared with other existing studies. We adopted the following two approaches as baselines. One (First Sense Only, FSO) is to immediately determine the first sense of the target word as listed in the dictionary. This is similar to the Most Frequent Sense method, which has also been used as the baseline in the existing supervised WSD studies. In such manner, the sense of the target word is basically assigned to the one with the highest frequency in the corpus. In the dictionary, each word contains all the possible senses, listed in the order in which they are frequently used in real life. In addition, the number of example sentences describing the meaning also depends largely on the order in which they are listed. The other baseline approach (C2V+k-NN) is a neural language model called context2vec that represents context embeddings describing the contextual information of the word in the context. It adopts the bidirectional LSTM to obtain the contextual embedding for the target word and assign the sense most similar to the context information of the target word in the training data using the k-nearest neighbor algorithm. Therefore, its performance has been shown quite dependent on that of the neural language model.

Table 3 demonstrates that our proposed system outperformed the two comparison models over our Oxford Dictionary based WSD dataset. It implies that the way of individually predicting meanings according to the parts of speech of words has shown its feasibility on data with high levels of sense ambiguity. Obviously seen from the comparison with the FSO method, the semantic analysis of words in the context is much more accurate on WSD than the simply matching method. In addition, our proposed model has shown 3.7% higher accuracy than that of the context2vec which determines the senses of the words solely based on the similarities of contextual embeddings. Since both approaches are given a similar way of obtaining contextual embeddings of the target word from the bi-directional LSTM, it is reasonable to highlight that the performance difference stems from our part-of-speech based sense classification method.

TABLE 3. Model performance for the Oxford dataset.

Method	Accuracy (%)	Noun (%)	Verb (%)	ADJ (%)	Adverb (%)
FSO	44.38	44.1	41.0	50.0	42.7
C2V+kNN	61.05	61.4	59.7	65.1	60.0
Our model	**64.75**	**64.3**	**62.1**	**70.5**	61.9
w/o hybrid	63.01	63.7	60.5	66.7	62.1

Next, we also conducted an ablation study to verify our hybrid sense classification method. Through various experiments, we obtained the optimal performance as shown in Table 3 when we set the number of frequently used meanings to 4. By adding the hybrid method, it has the effect of increasing accuracy by approximately 1.7%. We further analyzed the accuracy of the evaluation data by dividing them in the order of pre-listing for each meaning, as shown in Figure 3. The X axis refers to the order of the sense listing with the number of the example sentences. The Y axis represents the accuracy. As shown in the graph, the number of the example sentences of the senses varies considerably according to the order of senses in the dictionary. This distribution is also the same as the training set. Therefore, although the overall tendency of accuracy decreases with respect to the order of senses in the dictionary, it can be seen that the hybrid prediction method contributes to the improvement of classification accuracy. In particular, the improved accuracy of the rare senses is more pronounced. Since there is a significant variation in the number of the example sentences for each sense, the context feature still contains errors from the neural language model. Nonetheless, the result can be compensated

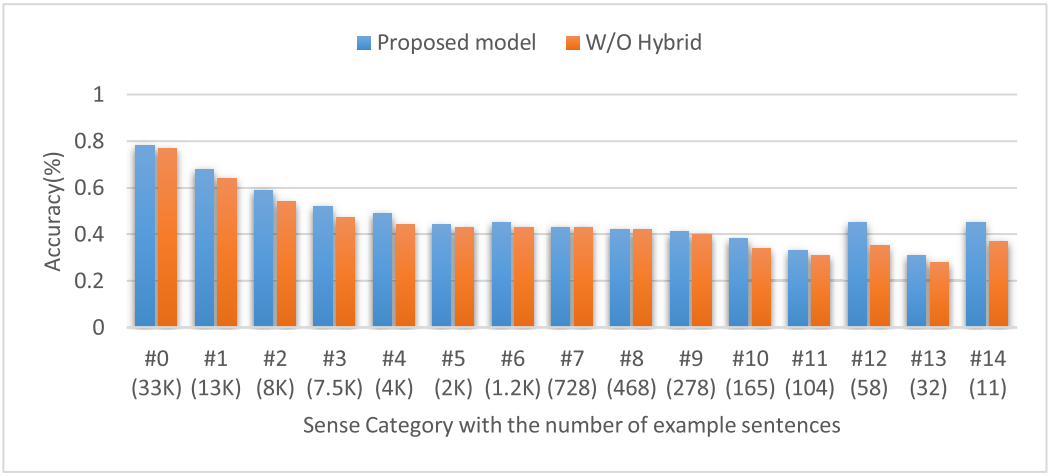

FIGURE 3. Effects of hybrid sense classification.

TABLE 4. Examples with positive results.

Example 1	Sentence	The ruling Democratic Party of Korea wants to start **deliberating** Liberty opposition main the but, impact maximum for possible as soon as...
	Predicted Sentence (Correct)	engage in long and careful consideration
	First Sense	done consciously and intentionally
Example 2	Sentence	He said this week's national **strikes** in, country the of areas all in demonstrate to going are...
	Predicted Sentence (Correct)	a refusal to work organized by a body of employees as a form of protest, typically in an attempt to gain a concession or concessions from their employer
	First Sense	hit forcibly and deliberately with one's hand or a weapon or other implement

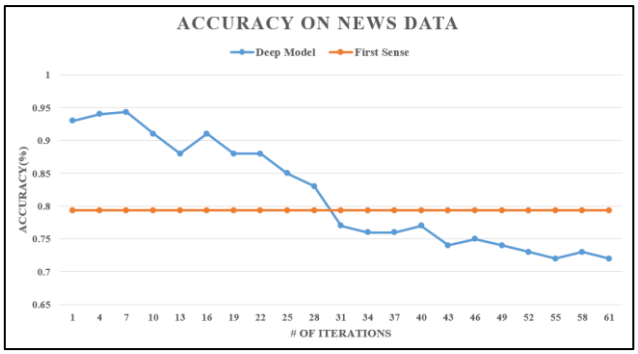

FIGURE 4. Accuracy of news article.

by separately predicting rare senses. Therefore, the hybrid sense classification method plays a role in supplementing the performance for less-frequently used senses.

Furthermore, we conducted another experiment to investigate the performance of the proposed model over a more realistic environment. As seen in Figure 4, First Sense denotes the output of the very first sense ID that appears in the Oxford Dictionary, and *Deep Model* is the output of the sense ID in accordance with the approach proposed in this study. The y-axis refers to accuracy and x-axis to the number of learning iterations. First Sense in the News Test showed an accuracy of 79.36%. The number was relatively higher as the usage frequency in real life is implemented less in the dictionary. Although the suggested model achieved an accuracy of 94.33% after seven learning iterations, the performance deteriorated as the number of iterations increased. This issue is due to different sentence patterns and portions of sense ID in the Oxford Test and News Test. Hence, with more learning iterations, the model parameters are over-fit to the Oxford Training corpus, causing performance degradation in the case of news. Therefore, adequate tuning is necessary to implement a WSD model.

We conducted exhaustive analysis on our experiment results. Table 4 demonstrates how the suggested model offers an accurate sense analysis of the target word in a random sentence. To sum, the proposed model can be an effective

solution to the second language students or translators when recognizing the accurate sense of the word in the context.

VI. CONCLUSION

This paper proposed a novel architecture for WSD capable of addressing a versatile large-scale dataset. To show its capacity, we introduced a new dataset for WSD that was automatically constructed from the certified and highly reliable Oxford Dictionary. Because the dictionary contains several words with a wide range of domains and all the known senses, it is normally considered to be one of the standard sources for sense identification. Though existing studies have shown considerable success, there still remains a threshold because they only cover a limited number of domains due to the limited nature of the training data. In contrast, our proposed model can address texts with a wide range of domains, which expands its capacity and enables success in various applications such as second language education, medicine, and more.

In the future, we will extend our work to incorporate multimodal data such as images or audios. This should ultimately allow us to extend the capacity of our model to utilize content with various formats.

REFERENCES

[1] H. Ahn, M. Seo, C. Park, J. Kim, and J. Seo, "Extensive use of morpheme features in korean dependency parsing," in *Proc. IEEE Int. Conf. Big Data Smart Comput.*, Feb. 2019, pp. 1–4.

[2] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.

[3] A. M. Butnaru and R. T. Ionescu, "ShotgunWSD 2.0: An improved algorithm for global word sense disambiguation," *IEEE Access*, vol. 7, pp. 120961–120975, 2019.

[4] Y. S. Chan, H. T. Ng, and D. Chiang, "Word sense disambiguation improves statistical machine translation," in *Proc. 45th Annu. Meeting Assoc. Comput. Linguistics, Assoc. Comput. Linguistics*, Prague, Czech Republic, 2007, pp. 33–40.

[5] D. S. Chaplot, P. Bhattacharyya, and A. Paranjape, "Unsupervised word sense disambiguation using Markov random field and dependency parser," in *Proc. 29th AAAI Conf. Artif. Intell. (AAAI)*, 2015, pp. 2217–2223.

[6] D. Chen and C. Manning, "A fast and accurate dependency parser using neural networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 740–750.

[7] H.-Y. Chen and K.-Y. Liu, "Web-based synchronized multimedia lecture system design for teaching/learning Chinese as second language," *Comput. Edu.*, vol. 50, no. 3, pp. 693–702, Apr. 2008.

[8] T. Chklovski and R. Mihalcea, "Building a sense tagged corpus with open mind word expert," in *Proc. Workshop Word Sense Disambiguation Recent Successes Future Directions*, 2002, pp. 116–122.

[9] R. Collobert, "Deep learning for efficient discriminative parsing," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, Fort Lauderdale, FL, USA, 2011, pp. 224–232.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[11] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, Fort Lauderdale, FL, USA, 2011, pp. 315–323.

[12] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013. pp. 6645–6649.

[13] T. He, J. Zhang, Z. Zhou, and J. Glass, "Quantifying exposure bias for neural language generation," 2019, *arXiv:1905.10617*. [Online]. Available: http://arxiv.org/abs/1905.10617

[14] Y. Heo, S. Kang, and D. Yoo, "Multimodal neural machine translation with weakly labeled images," *IEEE Access*, vol. 7, pp. 54042–54053, 2019.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.

[16] I. Iacobacci, M. T. Pilehvar, and R. Navigli, "Embeddings for word sense disambiguation: An evaluation study," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics, Assoc. Comput. Linguistics*, Berlin, Germany, 2016, pp. 897–907.

[17] S. Jaf and C. Calder, "Deep learning for natural language parsing," *IEEE Access*, vol. 7, pp. 131363–131373, 2019.

[18] S.-Z. Lee, J.-I. Tsujii, and H.-C. Rim, "Hidden Markov model-based Korean part-of-speech tagging considering high agglutinativity, word-spacing, and lexical correlativity," in *Proc. 38th Annu. Meeting Assoc. Comput. Linguistics*, Hong Kong, 2000, pp. 384–391.

[19] D. Loureiro and A. Jorge, "Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 5682–5691.

[20] M.-T. Luong and C. D. Manning, "Achieving open vocabulary neural machine translation with hybrid word-character models," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, 2016, pp. 1054–1063.

[21] J. Ma, Y. Zhang, and J. Zhu, "Tagging the Web: Building a robust Web tagger with neural network," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, Baltimore, MA, USA, 2014, pp. 144–154.

[22] J. Ma, J. Zhu, T. Xiao, and N. Yang, "Easy-first pos tagging and dependency parsing with beam search," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics, Assoc. Comput. Linguistics*, Sofia, Bulgaria, 2013, pp. 110–114.

[23] X. Ma, Z. Hu, J. Liu, N. Peng, G. Neubig, and E. Hovy, "Stack-pointer networks for dependency parsing," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1403–1414.

[24] O. Melamud, J. Goldberger, and I. Dagan, "Context2vec: Learning generic context embedding with bidirectional LSTM," in *Proc. 20th SIGNLL Conf. Comput. Natural Lang. Learn.*, Berlin, Germany, 2016, pp. 51–61.

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[26] G. A. Miller, "WordNet: A lexical database for english," *Commun. ACM*, vol. 38, pp. 39–41, Nov. 1995.

[27] G. A. Miller, M. Chodorow, S. Landes, C. Leacock, and R. G. Thomas, "Using a semantic concordance for sense identification," in *Proc. Workshop Hum. Lang. Technol. (HLT)*, 1994, pp. 240–243.

[28] R. Navigli and M. Lapata, "Graph connectivity measures for unsupervised word sense disambiguation," in *Proc. 20th Int. Joint Conf. Artif. Intell.* San Francisco, CA, USA: Morgan Kaufmann, 2007, pp. 1683–1688.

[29] R. Navigli and S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artif. Intell.*, vol. 193, pp. 217–250, Dec. 2012.

[30] Q.-P. Nguyen, A.-D. Vo, J.-C. Shin, P. Tran, and C.-Y. Ock, "Korean-vietnamese neural machine translation system with korean morphological analysis and word sense disambiguation," *IEEE Access*, vol. 7, pp. 32602–32616, 2019.

[31] P. Norvig, "Inference in text understanding," in *Proc. 6th Nat. Conf. Artif. Intell.*, 1987, pp. 561–565.

[32] O. Dongsuk, S. Kwon, K. Kim, and Y. Ko, "Word sense disambiguation based on word similarity calculation using word vector representation from a knowledge-based graph," in *Proc. 27th Int. Conf. Comput. Linguistics, Assoc. Comput. Linguistics*, Santa Fe, NM, USA, 2018, pp. 2704–2714.

[33] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1532–1543.

[34] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, New Orleans, LA, USA, 2018, pp. 2227–2237.

[35] A. Raganato, J. Camacho-Collados, and R. Navigli, "Word sense disambiguation: A unified evaluation framework and empirical comparison," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics, Assoc. Comput. Linguistics*, Valencia, Spain, 2017, pp. 99–110.

[36] A. Raganato, C. Delli Bovi, and R. Navigli, "Neural sequence learning models for word sense disambiguation," in *Proc. Conf. Empirical Methods Natural Lang. Process., Assoc. Comput. Linguistics*, Copenhagen, Denmark, 2017, pp. 1156–1167.

[37] A. Rios Gonzales, L. Mascarell, and R. Sennrich, "Improving word sense disambiguation in neural machine translation with sense embeddings," in *Proc. 2nd Conf. Mach. Transl., Assoc. Comput. Linguistics*, Copenhagen, Denmark, 2017, pp. 11–19.
[38] F. Schmidt, "Generalization in generation: A closer look at exposure bias," in *Proc. 3rd Workshop Neural Gener. Transl., Assoc. Comput. Linguistics*, Hong Kong, 2019, pp. 157–167.
[39] H. Shen, R. Bunescu, and R. Mihalcea, "Coarse to fine grained sense disambiguation in wikipedia," in *Proc. 2nd Joint Conf. Lexical Comput. Semantics, Assoc. Comput. Linguistics*, Atlanta, GA, USA, 2013, pp. 22–31.
[40] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, "Parsing with compositional vector grammars," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics, Assoc. Comput. Linguistics*, Sofia, Bulgaria, 2013, pp. 455–465.
[41] C. Stokoe, M. P. Oakes, and J. Tait, "Word sense disambiguation in information retrieval revisited," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, New York, NY, USA, 2003, pp. 159–166.
[42] K. Stratos, M. Collins, and D. Hsu, "Unsupervised part-of-speech tagging with anchor hidden Markov models," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 245–257, Jun. 2016.
[43] K. Taghipour and H. T. Ng, "One million sense-tagged instances for word sense disambiguation and induction," in *Proc. 19th Conf. Comput. Natural Lang. Learn., Assoc. Comput. Linguistics*, Beijing, China, 2015, pp. 338–344.
[44] S. Taneja, C. Gupta, K. Goyal, and D. Gureja, "An enhanced K-nearest neighbor algorithm using information gain and clustering," in *Proc. 14th Int. Conf. Adv. Comput. Commun. Technol.*, 2014, pp. 325–329.
[45] H. Wang, Q. Zhang, and J. Yuan, "Semantically enhanced medical information retrieval system: A tensor factorization based approach," *IEEE Access*, vol. 5, pp. 7584–7593, 2017.
[46] W. Wang and B. Chang, "Graph-based dependency parsing with bidirectional LSTM," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 2306–2315.
[47] D. Weissenborn, L. Hennig, F. Xu, and H. Uszkoreit, "Multi-objective optimization for the joint disambiguation of nouns and named entities," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics*, Beijing, China, 2015, pp. 596–605.
[48] Z. Zhong and H. T. Ng, "It makes sense: A wide-coverage word sense disambiguation system for free text," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, Uppsala, Sweden, 2010, pp. 78–83.
[49] Z. Zhong and H. T. Ng, "Word sense disambiguation improves information retrieval," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics, Assoc. Comput. Linguistics*, Jeju Island, South Korea, 2012, pp. 273–282.
[50] Y. Zhonggen, "Differences in serious game-aided and traditional English vocabulary acquisition," *Comput. Educ.*, vol. 127, pp. 214–232, Dec. 2018.

YOONSEOK HEO received the B.S. and M.S. degrees in computer science (major in in natural language generation) from Sogang Unversity. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Sogang University. He worked as a Researcher with Gachon University, in 2018. He is interested in spoken dialogue system, machine translation, question answering, machine reading comprehension, and named entity recognition. His current research focuses on the way of exploiting multimodal resources for machine translation and addressing large-scale open domain texts for machine reading comprehension.

SANGWOO KANG received the Ph.D. degree in computer science from Sogang University. He was a Research Fellow Professor with Sogang University. He has been an Assistant Professor with the Department of Software, Gachon University, since September 2016. He is currently leading the Intelligent Software and Natural Language Processing Laboratory, Gachon University. His specialty is natural language processing, and he is interested in spoken dialogue interface, information retrieval, text mining, opinion mining, big data, and UI/UX. His recent focus has been in applying deep learning techniques to his research.

JUNGYUN SEO received the B.S. degree in mathematics and the M.S. and Ph.D. degrees in computer science from the Department of Computer Science, The University of Texas at Austin, Austin, in 1981, 1985, and 1990, respectively.

Since 1991, he returned to join the Faculty of Korea Advanced Institute of Science and Technology, Taejon, where he leaded the Natural Language Processing Laboratory, Computer Science Department. In 1995, he moved to Sogang University, Seoul, and became a Full Professor, in 2001. He serves as the President of Korea Information Science Society, in 2013. His research interests include multimodal dialogues, statistical methods for NLP, machine translation, and information retrieval.