Corpus-Aware Graph Aggregation Network for Sequence Labeling

Haibin Chen, Qianli Ma, *Member, IEEE*, Liuhong Yu, Zhenxi Lin, and Jiangyue Yan

***Abstract*—Current state-of-the-art sequence labeling models are typically based on sequential architecture such as Bi-directional LSTM (BiLSTM). However, the structure of processing a word at a time based on the sequential order restricts the full utilization of non-sequential features, including syntactic relationships, word co-occurrence relations, and document topics. They can be regarded as the corpus-level features and critical for sequence labeling. In this paper, we propose a Corpus-Aware Graph Aggregation Network. Specifically, we build three types of graphs, i.e., a word-topic graph, a word co-occurrence graph, and a word syntactic dependency graph, to express different kinds of corpus-level non-sequential features. After that, a graph convolutional network (GCN) is adapted to model the relations between words and non-sequential features. Finally, we employ a label-aware attention mechanism to aggregate corpus-aware non-sequential features and sequential ones for sequence labeling. The experimental results on four sequence labeling tasks (named entity recognition, chunking, multilingual sequence labeling, and target-based sentiment analysis) show that our model achieves state-of-the-art performance.**

***Index Terms*—Corpus-aware features, graph neural network (GNN), attention mechanism, sequence labeling.**

I. Introduction

SEQUENCE labeling tasks are fundamental problems of Natural Language Processing (NLP), aiming to classify words in a sentence into several predefined speech or entity tags. Recently, BiLSTMs have reached a remarkable performance [1]–[3] and have become a dominant method due to its powerful ability on sequence modeling. However, such sequential methods take the current word as input and accumulate the hidden states sequentially, making it challenging to capture non-sequential dependencies [4], [5]. Some approach [6] has shown non-sequential information is highly useful for modeling sequence. In general, non-sequential features can be classified into sentence-level non-sequential features (e.g., the relation between words of a sentence) and corpus-level non-sequential features (e.g., the document topic of the whole corpus).

Existing methods mainly focus on modeling sentence-level non-sequential dependencies. To capture sentence-level non-sequential features from the whole sentence, some researchers tried to introduced non-sequential dependencies through dependency trees [7], or deepen the state transition by a deep transition architecture [4]. They followed the traditional way and considered only one sentence at a time. However, they ignored corpus-level relations such as the latent topic representation and word co-occurrence. Their methods prevent the model from capturing corpus-level global information, which leads to restricted performance [8].

In addition, some recent studies introduced sentence-level non-sequential features through leveraging information from the document. Existing document-aware work such as [5] propagated word occurrence information in a document through graph convolution network. However, compared with sentence-level methods, the nature of document-aware methods still has not solved some difficulties such as OOV (out-of-vocabulary) problem, because it is challenging to identify unseen words through few occurrences information [9]. The meaning of words, especially for words that are difficult to understand, needs to be inferred by combining sentence-level semantics and corpus-level semantics. More specifically, different occurrences of a particular context are very likely to have the same labeling types in the corpus [10]. So the word labeling types can be inferred from these common expressions, which can be reflected by corpus-level word co-occurrence relations. Moreover, sequence labeling types are related to its corpus summarization [11], which can be reflected by corpus-level topic modeling. From the perspective of the whole corpus instead of several sentences, richer semantics like corpus-level co-occurrence information and corpus topic semantics are helpful to enhance the word representations especially for the OOV word.

For example, in Figure 1, entity word "X-box" is an out-of-vocabulary word in pre-trained word embedding Glove [12]. The baseline sentence-level methods Bi-LSTM+CRF incorrectly marks "X-box" as type ORGANIZATION, probably because sentence-level sequential information including the organization words "Microsoft" and "Nintendo" misleads the model. In stead of considering one sentence at a time, the challenge is to infer the types of difficult words from the corpus topic and corpus word usage. In this example we find that the key words "Microsoft, Nintendo, machine..." are highly relevant to the topic of

Manuscript received October 26, 2020; revised February 26, 2021; accepted May 10, 2021. Date of publication May 26, 2021; date of current version June 14, 2021. This work was supported in part by the National Natural Science Foundation of China under Grants 61502174 and 61872148, by the Natural Science Foundation of Guangdong Province under Grants 2017A030313355, 2019A1515010768, and 2021A1515011496, by Guangzhou Science and Technology Planning Project under Grants 201704030051 and 201902010020, by the Key R&D Program of Guangdong Province under Grant 2018B010107002, and by the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Taro Watanabe. *(Corresponding author: Qianli Ma.)*

The authors are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: haibin_chen@foxmail.com; qianlima@scut.edu.cn; yu.liuhong@foxmail.com; zhenxi_lin@foxmail.com; jiangyue9606@gmail.com).

Digital Object Identifier 10.1109/TASLP.2021.3084105

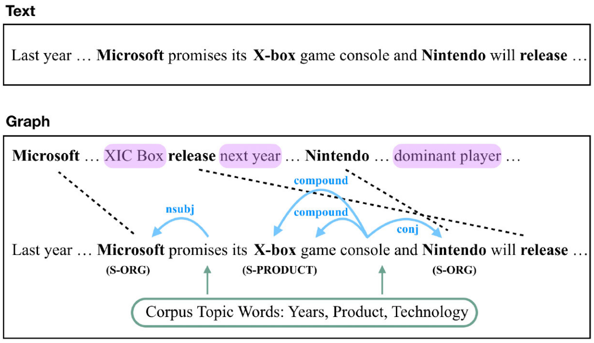

Fig. 1. Example of the entity recognition task with an out-of-vocabulary word "X-box". Corpus-level word dependencies (purple windows, connect with other document by dotted lines), corpus-level topic semantics (green arrows) and sentence-level syntactic dependencies (blue arrows) provide non-sequential valuable information to recognize an unseen word "X-box" as entity type Product.

game and product. From the topic model, topic words "years, product and technology" can also help to infer the word meanings. Moreover, corpus-level word co-occurrence information, which aggregates local contextual information from the corpus is beneficial to infer the word meanings. Besides, sentence-level dependencies show that "X-box game" is a compound noun modifying word "console," while organization words are usually marked as proper nouns. By extracting these non-sequential features, words with similar meanings, words collocation and grammatical structure can be fully utilized to solve the challenge of modeling corpus-level features. The example in Figure 1 shows that corpus-level word dependencies, corpus-level topic semantics and sentence-level syntactic dependencies are helpful to recognize sequence labels.

This paper proposes a corpus-aware graph aggregation network (CAGAN) and learns non-sequential features from both corpus-level and sentence-level. In particular, we construct three graphs in CAGAN, i.e., word-topic graph, word co-occurrence graph, and sentence dependency graph, to model three types of non-sequential dependencies. Concretely, we utilize neural topic model LDA [13] to model soft sentence topic information and build a word-topic graph by connecting virtual topic nodes and words in the input sentence. Based on the word-topic graph, we can capture the relationship between words and topics. We also build a word co-occurrence graph based on the words' co-occurrence relationship, where the information of similar words and common grammatical structures can be learned. Simultaneously, for non-sequential information in the sentence, we build a dependency graph to model syntactic relations. After that, a graph convolution network is employed to aggregate the different learned non-sequential features. Furthermore, we proposed a label-aware attention mechanism to fuse the three kinds of non-sequential features and original sequential ones. Therefore, CAGAN can model not only sequential features but also three different types of non-sequential dependencies. Experimental results on named entity recognition, syntactic chunking, multilingual POS tagging, and target-based sentiment analysis demonstrate that the corpus-aware non-sequential representations are useful to sequence labeling tasks.

The main differences between our model, pool-Flair [14], and Cross-sentence Context [15] are: First, pool-Flair introduced an extra character-based language model, but our model and Cross-sentence Context do not need to pre-train an extra character-level language model. Second, pool-Flair employed external memory to save the word embedding that the model encounters, while our model connects the related words by heterogeneous graphs directly. Third, constructing multiple input sentences to obtain document-level information [15] may be limited by BERT's maximum length and excessive memory. Therefore, our model has less memory requirement and performs sequence-level labeling awareness of corpus-level information.

Our contributions are summarized as follows.

- To tackle the challenge of modeling corpus-level features, we are the first to introduce corpus-level non-sequential features including latent topic information and word co-occurrence information for sequence labeling.
- Moreover, we propose corpus-aware graph aggregation network to dynamically model non-sequential features and fuse them with label-aware attention mechanism.
- Our model outperforms previous systems on different subtasks of sequence labeling. Furthermore, we conduct elaborate analysis of corpus-level non-sequential features.

II. Related Work

A. Sequence Labeling

Recent advances in deep neural models allow us to build reliable NER systems without hand-crafted features. Sequential architectures have become a dominant method for sequence modeling. BiLSTM is a commonly used model designed to prevent vanishing gradient problem. [1] leveraged CRF on the output layer of BiLSTM (BiLSTM-CRF) to extract sequential features, which considered adjacent labels when predicting labels. After that, BiLSTM-CRF was further improved by character-level layer based on BiLSTM and CNN [2], [3]. Other improved methods like [16] introduced additional segment-level LSTM to model segmentation information, which is not explicitly modeled by sequential models. These variations of BiLSTM mainly focus on extracting sequential features. More recently, language models trained on massive corpora in both character level (ELMo [17], Flair [14]) and token level (BERT [18]) greatly enhanced the representation of words. In this work, we conduct experiments on both conditions with/without ELMo.

B. Non-Sequential Features Learning

Recent advances in sequence modeling have shown the usefulness of non-sequential features [6]. Existing methods on sequential labeling mainly focus on sentence-level non-sequential features and extract the word relations, which are not in sequential order. GCDT [4] investigated a Global Context-enhanced Deep Transition architecture by averaging the sentence-level hidden states of each word from an independent deep transition RNN. Jie *et al.* [7] proposed a dependency-guided LSTM-CRF model to encode the complete dependency trees. However, a limitation of these models is that they only consider the information of

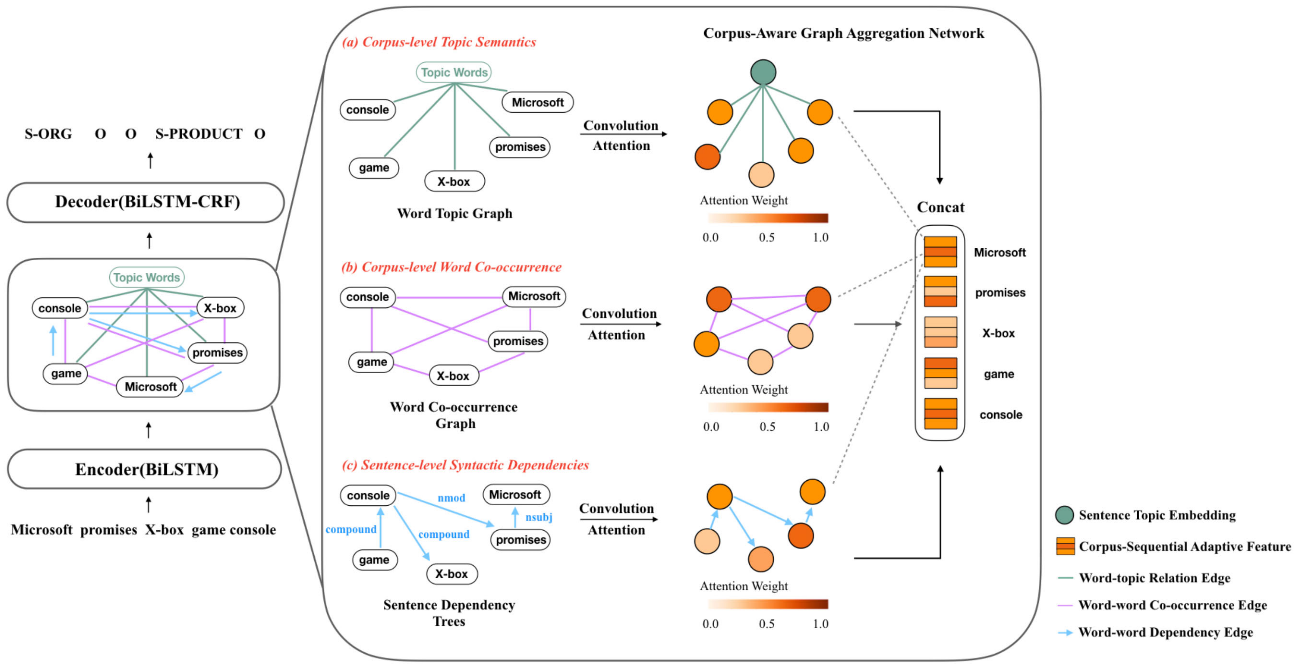

Fig. 2. The overall architecture of the proposed model. The encoder (BiLSTM) extracts the sequential feature. Then, the model builds three kinds of non-sequential graph Word Topic Graph, Word Co-occurrence Graph and Sentence Dependency Trees. Non-sequential features are extracted by graph convolution and attention mechanism. Different color nodes represent the attention weights of different non-sequential feature. Finally, these features are concatenated and fed into decoder (BiLSTM-CRF) to perform sequence labeling.

one sentence at a time without considering the document or corpus information, which leads to restricted performance [8]. Recent document-aware work such as Luo *et al.* [8] proposed a key-value memory network to memorize all the document-level features and designed an attention-based distillation approach, which required large memory. GraphIE [5] used graph convolution network to propagate word occurrence information between non-local words.

However, the perspective of document-aware still faces some difficulties in recognizing some challenging words such as OOV (out-of-vocabulary) words. The challenge is how to fully utilize corpus-level features, and infer the word meanings through corpus topic, words collocation and grammatical structure. Corpus-level means we taking all documents into account, not a single document composed of several sentences.

III. Proposed Model

In this section, we will describe the details about our corpus-aware graph aggregation network. Figure 2 shows the overall architecture of our proposed model.

A. Encoder

In the sequence labeling task, given the input sequence $x_{seq} = \{x_1, \ldots, x_t, \ldots, x_n\}$, the model will predict the label sequence $y_{seq} = \{y_1, \ldots, y_t, \ldots, y_n\}$ where n is the number of words. Our model first captures each token representation x_t by concatenating character level word embedding c_t obtained from BiLSTM and pre-trained word embedding w_t obtained from Glove. Then the input $x_t = [w_t; c_t]$ is fed to encoder BiLSTM to extract sequential feature for each word.

$$H_t^{sequential} = [\overrightarrow{h_t}; \overleftarrow{h_t}], \tag{1}$$

$$\overrightarrow{h_t} = LSTM(x_t; \overrightarrow{h}_{t-1}), \tag{2}$$

$$\overleftarrow{h_t} = LSTM(x_t; \overleftarrow{h}_{t-1}), \tag{3}$$

where $[;]$ is the concatenation operation. $h_t \in \mathbb{R}^{n\times(2\times d_{lstm})}$ is the contextual representation of the t-th word by concatenating the forward hidden state $\overrightarrow{h}_t$ and the backward hidden state $\overleftarrow{h}_t$.

B. Corpus-Aware Graph Aggregation Network

In the following subsection, we will introduce three different non-sequential graphs, word-topic graph, word co-occurrence graph and sentence dependency graph, respectively. After that, we propose a label-aware attention mechanism to combine non-sequential and sequential features dynamically.

1) Definition: Due to the heterogeneity of topic information, co-occurrence information and syntactic information, we divide these non-sequential information into three homogeneous graphs to avoid their interference. For word-topic graph, we define $G^{topic} = (V^{words}, V^{topic}, E^{topic})$, where V^{topic} consists of the virtual topic nodes. E^{topic} are the undirected edges between words and topics. For word co-occurrence graph, we define $G^{occur} = (V^{words}, E^{occur})$, where E^{occur} are the undirected co-occurrence information. For sentence dependency graph, we define $G^{depend} = (V^{words}, E^{depend})$, where E^{depend} are the directed edges of syntactic dependencies. The initial representations of V^{words} are sequential features h_t from the encoder.

2) Word-Topic Graph: We employ latent topics of the document to obtain the corpus-level topic semantic representation and document abstraction. For $G^{topic} = (V^{words}, V^{topic}, E^{topic})$. V^{topic}, E^{topic} are obtained from topic models. In the following paragraphs, we will introduce the extraction of corpus-aware topic information and the construction of the graph.

Firstly, we mine the latent topics using LDA [13]. Each topic distribution is represented by a probability distribution over the words. More specifically, $t_i = (\theta_1, \ldots, \theta_w)(t_i = \{t_1, \ldots, t_i, \ldots, t_k\}$, k denotes the number of topics, w denotes the vocabulary size). We extract the top p words for each topic distribution, and then denote average pooling of these top p words embeddings as topic representations $V^{topic} = (V_1, \ldots, V_k), V^{topic} \in \mathbb{R}^{p \times d_{word}}$, d_{word} is the dimension of pre-trained word embedding. The edge weights $E^{topic} \in \mathbb{R}^{n \times k}$ are the normalized topic probabilities(obtained from topic models) between the corresponding corpus and every topics.

Based on the nodes and edges above, adjacency matrix $A^{topic} \in \mathbb{R}^{(k+n)\times(k+n)}$ is defined by connecting every word with k virtual topic nodes. The corresponding degree matrix M^{topic} is calculated by $M_{ii} = \sum_j A_{ij}$. The word topic graph models the corpus-level topic semantics through the graph aggregation between words and topics. The aggregation rule for each layer l is formulated as follows:

$$H^{topic}_{(l+1)} = ReLU\left(\tilde{A}^{topic} \cdot H^{topic}_{(l)} \cdot W^{topic}_{(l)}\right), \tag{4}$$

where $\tilde{A}^{topic} = (M^{topic})^{-\frac{1}{2}} A^{topic} (M^{topic})^{-\frac{1}{2}}$ is the symmetric normalized adjacency matrix. $W^{topic}_{(l)} \in \mathbb{R}^{2\times d_{lstm} \times d_{topic}}$ is the weight matrix for layer l, where d_{topic} is the output dimension of word topic graph. $H^{topic} \in \mathbb{R}^{n \times d_{topic}}$ is the output of word topic graph. In Figure 2 (a) corpus-level topic semantics, the OOV word "X-box" interacts with virtual topic nodes (largest probability topic words "years, product, technology") and obtains higher-order topic semantics.

3) Word Co-Occurrence Graph: Co-occurrence information can provide valuable non-sequential word dependencies such as similar meaning words and word collocations. We employ pointwise mutual information (PMI) motivated by [19], which is a popular method to measure the word associations. A fixed-size sliding window and a minimum frequency threshold are set to preprocess the dataset and get the corpus-level PMI. The PMI value of a word pair i, j is computed as follows:

$$\mathrm{PMI}(i,j) = \log \frac{p(i,j)}{p(i)p(j)}, \tag{5}$$

$$p(i,j) = \frac{\#W(i,j)}{\#W}, \tag{6}$$

$$p(i) = \frac{\#W(i)}{\#W}, \tag{7}$$

where $\#W(i)$ is the number of sliding windows in a corpus that contain word i, $\#W(i,j)$ is the number of sliding windows that contain both word i and j, and $\#W$ is the total number of sliding windows in the corpus. A positive PMI value represents a high non-sequential correlation of word pairs.

For $G^{occur} = (V^{words}, E^{occur})$, the PMI of every word pair is the edge weight of $E^{occur} \in \mathbb{R}^{n\times n}$. Then, we let $A^{occur} \in \mathbb{R}^{n\times n}$ be an adjacency matrix by connecting every two words. Similarly, M^{occur} is a degree matrix calculating by $M_{ii} = \sum_j A_{ij}$. And the aggregation rule for each layer l is formulated as follows:

$$H^{occur}_{(l+1)} = ReLU\left(\tilde{A}^{occur} \cdot H^{occur}_{(l)} \cdot W^{occur}_{(l)}\right), \tag{8}$$

where $\tilde{A}^{occur} = (M^{occur})^{-\frac{1}{2}} A^{occur} (M^{occur})^{-\frac{1}{2}}$ is the symmetric normalized adjacency matrix. $W^{occur}_{(l)} \in \mathbb{R}^{2\times d_{lstm} \times d_{occur}}$ is the weight matrix for layer l, where d_{occur} is the output dimension of word co-occurrence graph. $H^{occur} \in \mathbb{R}^{n \times d_{occur}}$ is the output of word co-occurrence graph. In Figure 2 (b) corpus-level word co-occurrence, the OOV words "X-box" enriched by the high co-association words "Microsoft, Nintendo, release".

4) Sentence Dependency Graph: Non-sequential syntactic relations and grammar structure can also infer potential speech tags. For datasets that do not provide dependency tags, we use an external parser SpaCy[1] as preprocessing. We decompose the directed dependency edge into the head index and edge type [20]. Head index and edge type share the same dependency graph, but update with the different transformation matrix. Then we sum these two kinds of features up to get the graph output H^{dep}. $A^{dep} \in \mathbb{R}^{n\times n}$ is the adjacency matrix representing the head index. The aggregation rule is formulated as follows:

$$B^{dep}_l = \left(W^{dep1}_{(l)} H^{dep}_{(l-1)} + W^{dep2}_{(l)} H^{dep}_{(l-1)} e_r\right), \tag{9}$$

$$H^{dep}_{(l+1)} = ReLU\left(\sum A^{dep} B^{dep}_{(l)}\right), \tag{10}$$

where $W^{dep1}_{(l)} \in \mathbb{R}^{2\times d_{lstm} \times d_{dep}}$ is the weight matrix of the head index and $W^{dep2}_{(l)} \in \mathbb{R}^{2\times d_{lstm} \times d_{dep}}$ is the weight matrix of edge type, where d_{dep} is the output dimension of sentence dependency graph. $e_r \in \mathbb{R}^n$ is the edge type embedding and initialize randomly during training. $H^{dep} \in \mathbb{R}^{n \times d_{dep}}$ is the output of sentence dependency graph. We do not gain significant improvement from multi-hop fusion or other gating mechanisms. In Figure 2 (c) sentence-level syntactic dependencies, "X-box game" is a compound noun modifying "console" instead of a proper noun like other organization words "Nintendo, Microsoft".

5) Label-Aware Attention Mechanism: Through three non-sequential graphs, the model extracts three non-sequential features. Words with different feature types benefit tagging from different aspects. For example, the phrase "X-box game console" contributes more to grammar structure and words "Microsoft, release" contribute more to word co-occurrence. Therefore, it is necessary to consider the fusion way of three non-sequential features and original sequential features. We propose an attention scorer function for modeling type importance. Given graph network output $H \in \mathbb{R}^{n\times(d_{topic}+d_{occur}+d_{dep})}$, we design

[1][Online]. Available: https://spacy.io/

a type-aware scorer function $score$ as follows:

$$H = [H^{topic}; H^{occur}; H^{dep}], \tag{11}$$

$$score(H) = sigmoid(W^{att} \odot Pool(H) + b^{att}), \tag{12}$$

where $W^{att} \in \mathbb{R}^{3\times n}$ is transformation matrix and $b^{att} \in \mathbb{R}^3$ is bia term. $Pool()$ denotes that we use average pooling to represent different type features. Then we utilize scorer function weight to obtain fusion representations:

$$H^{non-sequential} = H \odot score(H), \tag{13}$$

where $\odot$ denotes element-wise multiplication. $H^{non-sequential} \in \mathbb{R}^{n\times(d_{topic}+d_{occur}+d_{dep})}$ are the fused representations of non-sequential features.

Motivated by [21] and [8], we incorporate label information into feature fusion. The model attempt to make word representation closer to the label embedding during training(e.g. word representation of "Microsoft" close to label embedding "ORGANIZATION"). The word representations H^{total} are obtained by concatenating non-sequential and sequential features. Then, the cosine similarity between word representation and label embedding is adopted to measure the relation of every word-label pair. The relative word-label confidence scores $\beta_{label} \in \mathbb{R}^n$ are further obtained through a simple convolutional neural network(CNN)(the same as [8]). Finally, word-label confidence scores are weighted to word representation and get a label-aware word representation H^{merge}:

$$H^{total} = [H^{non-sequential}; H^{sequential}], \tag{14}$$

$$\beta_{label} = \text{CNN}(\text{cosine}(H^{total}, Label)), \tag{15}$$

$$H^{merge} = [\text{softmax}(\beta_{label}) \odot H^{total}; H^{total}], \tag{16}$$

where $\odot$ represents element-wise multiplication. The final hidden states H^{merge} are then fed to decoder to perform tagging.

C. Decoder

In our work, the decoder is instantiated as a BiLSTM+CRF tagger [3]. After Graph Aggregation Network, the corpus-sentence adaptive feature is propagated to each word through the BiLSTM. Finally, we use a conditional random field (CRF) classifier at the top of our model for sequential inference. CRF learns the transition rules from one label to another based on the feature vectors of input sequence as a whole instead of individually. Given the BiLSTM output S and target label sequence y_{seq}, the CRF layer defines the probability and computes the score of the entire sequence as follows:

$$S = BiLSTM(H^{merge}), \tag{17}$$

$$P(y|S) = \frac{exp(score(S), y))}{\sum_{y'} exp(score(S, y'))}, \tag{18}$$

$$score(S, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=0}^{n} F_{S, y_i}, \tag{19}$$

TABLE I
THE DATASET STATISTICS OF ONTONOTES 5.0, CONLL-2000, UD V2.2 ENGLISH AND CONLL-2003

Dataset		Train	Dev	Test
OntoNotes 5.0	Sent	59924	8528	8262
	Token	1088503	147724	152728
CoNLL-2000	Sent	8936	-	2012
	Token	211727	-	47377
UD v2.2 English	Sent	12544	2003	2078
	Token	204607	25120	25097
CoNLL-2003	Sent	14987	3466	3648
	Token	204567	51578	46666

TABLE II
THE OPINION TARGET STATISTICS OF LAPTOP, RESTAURANT AND TWITTER

Dataset		Train	Dev	Test	Total
Laptop	POS	883	104	339	1326
	NEG	754	106	130	990
	NEU	404	46	165	615
Restaurant	POS	2337	270	1524	4131
	NEG	942	93	500	1535
	NEU	614	50	263	927
Twitter	POS	-	-	-	692
	NEG	-	-	-	263
	NEU	-	-	-	2244

where $A_{y_i, y_{i+1}}$ is the transition parameter from label y_i to label y_{i+1}. F_{S, y_i} is an emission value representing the scores of label y_i of the i-th word.

We minimize the negative log-likelihood during training and viterbi algorithm is applied in searching for the label sequence during testing.

IV. EXPERIMENTS

A. Dataset

We verify our proposed model on four sequence labeling tasks including named entity recognition(OntoNotes 5.0 [22], CoNLL-2003 [23]), chunking(CoNLL 2000 [24]), multilingual POS tagging(Universal Dependencies(UD) v2.2 [25]) and target-based sentiment analysis(Laptop and Restaurant from from SemEval ABSA challenges [26]–[28]and Twitter from [29]). The OntoNotes 5.0 dataset consists of texts from a wide variety of sources, including newswire, magazine, Web text, broadcast conversation and so on. The CoNLL-2003 dataset includes four types of annotations: Person (PER), Location (LOC), Organization (ORG) and Miscellaneous (MISC). The CoNLL-2000 Chunking task defines 11 syntactic chunk types (NP, VP, PP, etc.). For multilingual POS tagging, we choose 8 resource-rich languages from UD v2.2 treebanks. OntoNotes 5.0 and Universal Dependencies datasets provide ground truth dependency labels. The definition of target-based sentiment analysis is detect both the boundary of target mention and the target sentiment as a sequence labeling task. Laptop and Restaurant are product review datasets. Twitter dataset will be conducted on ten-fold cross validation since there is no standard train-test split. We follow the standard dataset pre-process of Laptop, Restaurant and Twitter as done in [30]. Table I and Table II shows the statistics of dataset.

B. Experimental Setup

For word embeddings, we utilize the pre-trained $100D$ word embedding from Glove [12]. The character embedding is randomly initialized and the embedding size is 50. The LSTM hidden size is 200. For topic modeling, the topic number is 20. For word co-occurrence information extraction, the slide window size is 20 and the minimum frequency threshold is 3. The layer of graph aggregation network is 1. The output dimensions of word topic graph, word co-occurrence graph and dependency graph are $50, 100, 150$, respectively. Our model is optimized by SGD with a learning rate 0.01.

C. Comparison Models

For three sequence labeling tasks, the experimental results are compared with other start-of-the-art models such baseline BiLSTM-CRF [1], [3], GRN [31], dependency-guided LSTM-CRF [7], GCDT [4], BiLSTM-CRF with segment-level LSTM [16], LAN [32] and Flair [14]. The detail of the main comparison models used in paper are listed as follows:

BiLSTM-CRF [1], [3]: It is a hybrid bidirectional LSTM and CNN architecture to conbine word- and character-level features to conduct sequence labeling tasks.

GRN [31]: They propose a simple but effective CNN-based network gated relation network (GRN) for NER and the model also enjoys lower time costs to train and test.

Dependency-guided LSTM-CRF [7]: They encode the complete dependency trees and propose a dependency-guided LSTM-CRF model.

GCDT [4]: They deepen the state transition path in a sentence and obtain global representation for each word, and thus propose a Global Context enhanced Deep Transition architecture.

BiLSTM-CRF with segment-level LSTM [16]: They extend the baseline model with an additional segment-level layer called segLSTM, which employs an additional segment-level LSTM that trains features by learning adjacent context in a segment.

LAN [32]: They investigate a hierarchically-refined label attention network, which explicitly leverages label embeddings and captures potential long-term label dependency.

Flair [14]: They leverage the internal states of a trained character language model to obtain contextual string embeddings.

Hire-NER [8]: They extract sentence-level representation and document-level representation by label embedding attention network and memory network.

ACE [33]: They automate the process of finding better concatenations of different pre-trained embeddings.

HAST-TNet [34], [35]: HAST and TNet are the state-of-the-art models on the tasks of target boundary detection and target sentiment classification respectively. HAST-TNet is the pipline approach of these two models.

Li-unified [30]: They introduce two stacked recurrent neural networks and propose Boundary Guidance (BG) component, Sentiment Consistency (SC) component and Opinion-Enhanced (OE) Target Word Detection component.

D. Experimental Results

Table III, IV, VII, V and VI show the experimental results comparison between our proposed model and existing

TABLE III
THE EXPERIMENTAL RESULTS COMPARING TO OTHER STATE-OF-THE-ART MODELS ON ONTONOTES 5.0 DATASET

Models	F_1
BiLSTM-CRF (2016) [36]	86.28 ± 0.26
Shen-active (2017) [37]	86.63
Strubell-fast (2017) [38]	86.84 ± 0.19
GRN (2019) [31]	87.67 ± 0.17
Shin-segLSTM (2020) [16]	86.89 ± 0.11
Hire-NER (2020) [8]	87.98
Dependency-guided LSTM-CRF (2019) [7]	88.52
Ours	**88.75 ± 0.09**
+Language Models / External Knowledge	
Ghaddar-lexical (2018) [39]	87.95
Clark-semi (2018) [40]	88.88
Flair (2019) [14]	89.30
Dependency-guided LSTM-CRF (2019)(+ELMo) [7]	89.89
Hire-NER (2020)(+BERT) [8]	90.30
AESINER (2020)(+BERT) [41]	90.32
Ours+ELMo	89.94 ± 0.03
Ours+BERT	**91.32 ± 0.04**

TABLE IV
THE EXPERIMENTAL RESULTS COMPARING TO OTHER STATE-OF-THE-ART MODELS ON CONLL 2000 DATASET

Models	F_1
BiLSTM-CRF (2015) [1]	94.46
Zhai-neural (2017) [42]	94.72
Søgaard-multitask (2016) [43]	95.28
Xin-internal (2018) [44]	95.29 ± 0.08
Zhao-gated (2019) [45]	94.80 ± 0.05
GCDT (2019) [4]	95.25 ± 0.03[2]
Shin-segLSTM (2020) [16]	94.94 ± 0.06
Ours	**96.00 ± 0.05**
+Language Models / External Knowledge	
Liu-empower (2017) [46]	95.96 ± 0.08
Peters-semi (2017) [47]	96.37 ± 0.05
Flair (2018) [14]	96.72 ± 0.05
GCDT (2019)(+BERT) [4]	96.81 ± 0.03[2]
ACE (2020) [33]	97.00
Ours+ELMo	96.87 ± 0.03
Ours+BERT	**97.92 ± 0.02**

TABLE V
THE EXPERIMENTAL RESULTS COMPARING TO OTHER STATE-OF-THE-ART MODELS ON CONLL-2003 DATASET

Models	F_1
BiLSTM-CRF (2016) [3]	91.21
Strubell-fast (2017) [38]	90.54 ± 0.18
GRN (2019) [31]	91.44 ± 0.16
Zhao-gated (2019) [45]	91.26 ± 0.17
Shin-segLSTM (2020) [16]	91.48 ± 0.13
GCDT (2019) [4]	91.54[2]
MEID (2020) [48]	91.92
Hire-NER (2020) [8]	**91.96 ± 0.03**
Ours	91.61 ± 0.11
+Language Models / External Knowledge	
Liu-empower (2017) [46]	91.71 ± 0.10
Peters-semi (2017) [46]	91.93 ± 0.19
ELMo (2018) [17]	92.2
Shin-segLSTM (2020)(+ELMo) [16]	92.61
BERT (2018) [18]	92.80
GCDT (2019)(+BERT) [4]	93.23[2]
MEID (2020)(+Flair) [48]	93.09
Hire-NER (2020)(+BERT) [8]	93.37 ± 0.04
Cross-sentence (2020) [15]	93.44 ± 0.06
ACE (2020) [33]	**93.50**
Ours+ELMo	92.50 ± 0.05
Ours+BERT	93.21 ± 0.04

TABLE VI
THE EXPERIMENTAL RESULTS F1 ON THREE TARGET-BASED SENTIMENT ANALYSIS DATASET. * INDICATES OUR REIMPLEMENTATION RESULTS

Models	Laptop	Restaurant	Twitter
CRF-unified (2013) [29]	49.06	60.43	27.86
CRF-pipeline (2013) [29]	52.93	51.64	31.73
BiLSTM-CRF (2016) [3]	54.71	64.29	47.26
HAST-TNet (2018) [34], [35]	55.29	67.36	47.66
Li-unified (2019) [30]	57.90	69.80	48.01
BERT (2018) [49]	61.12	75.80*	53.81*
Ours	57.70	71.12	49.85
Ours+BERT	**64.21**	**76.25**	**55.20**

state-of-the-art methods on OntoNotes 5.0, CoNLL-2000, UD v2.2, CoNLL-2003, Laptop, Restaurant and Twitter dataset, respectively. Our model gives significant improvements over baseline BiLSTM-CRF on all datasets ($p < 0.01$). Comparing to the models that enhanced only by sentence-level non-sequential features [4], [7], our model achieves better results, which improves Dependency-guided LSTM-CRF [7] by 0.23 F1 on OntoNotes dataset, and improves GCDT [4] by 0.75 F1 on CoNLL-2000 dataset. Pre-trained external knowledge ELMo [17] and BERT [18] are adopted for comparison with the models which use the language models or external knowledge. Our proposed method also outperforms state-of-the-art methods such as Flair [14]. On UD v2.2 dataset, our model achieves state-of-the-art in 7 of 8 resource-rich language datasets, and improves 1.14 average F1 than LAN [32]. On CoNLL-2003 English dataset, our model outperforms baseline methods with or without pre-trained knowledge ELMo. The improvement is not significant compared with state-of-art methods. We found that introducing syntactic dependencies will harm the result of tagging, possibly due to the lower quality of dependency trees [7]. ACE [33] is a contemporary work and achieves the best results on CoNLL-2003. They incorporated 11 pre-trained embeddings to perform ensemble learning automatically, which inevitably has high memory requirements. They improved sequence labeling methods on embedding layers from a different perspective. On target-based sentiment analysis dataset, our model gets the similar result as [4] on Laptop dataset with GloVe embedding, and surpasses BERT's results on all datasets. The comparisons between our model incorporating with or without external knowledge and other state-of-the-art models demonstrate the usefulness of our proposed methods.

E. Ablation Analysis

"Baseline": Baseline BiLSTM-CRF.

"w/o Word Topic Graph": Not using the corpus-level topic semantics information.

"w/o Word Co-occurrence Graph": Not using the corpus-level word co-occurrence information.

"w/o Sentence Dependency Graph": Not using the sentence-level syntactic dependency information.

"w/o Graph Aggregation": Concatenating corpus-aware non-sequential features through MLP to sequential features (total parameters are the same as the graph aggregation network)

"w/o Label-aware Attention Mechanism": Not using the label-aware attention mechanism (concatenate all the features)

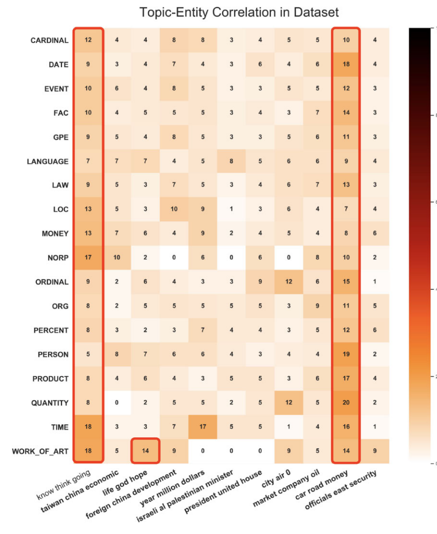

Fig. 3. Percentage of entity types (y axis) with respect to corresponding document topics (x axis) in the OntoNotes dataset.

We conduct an ablation study on our model without different corpus-aware non-sequential features and our modeling methods graph aggregation network. Table VIII shows that F1 will decrease with removing any corpus-aware non-sequential features. Without graph aggregation network, other modeling methods like concatenating corpus-aware non-sequential features through MLP to sequential features don't bring significant improvements. The attention mechanism is also beneficial for feature fusion.

F. Improvement Discussion

Table IX presents the F1 score of in-both-vocabulary words (IV), out-of-training-vocabulary words (OOTV), out-of-embedding-vocabulary words (OOEV), and out-of-both-vocabulary words (OOBV) on OntoNotes 5.0 dataset. Different corpus-aware information plays a different role in Out-of-Vocabulary (OOV) word performance. In IV and OOEV part, corpus-aware word co-occurrence information is helpful for the entities in the training set, which outperforms baseline by 1.40, 0.14 in terms of F1. Corpus-aware topic information is helpful for all the entities. There are 84.68% of the entities are located in the IV part, and the largest improvement comes from the IV part, which suggests that corpus-aware information fully utilizes the training corpus and brings significant improvements. These results suggest that corpus-aware information is powerful in recognizing entities.

G. Topic Visualization

To demonstrate whether the model benefits from the topic information, we first draw the heatmap in Figure 3 to show the relations between entity types and its corresponding document topic. We ignore the topic with percentage less than 10% for brevity, and sort the topic by frequency. Each topic is represented by the top 3 words with the largest probabilities(x-axis). Each line denotes the percentage of the entities with different topics.

TABLE VII
THE EXPERIMENTAL RESULTS COMPARING TO OTHER STATE-OF-THE-ART MODELS ON UD V2.2 DATASET, COMPARED ON 8 RESOURCE-RICH LANGUAGES

		cs	da	en	fr	nl	no	pt	sv	average
BiLSTM-softmax	mean	98.48	95.90	95.36	97.01	94.76	97.26	97.78	95.98	96.57
	± std	0.03	0.12	0.06	0.08	0.11	0.11	0.04	0.07	0.08
Yasunaga-Robust (2018) [50]	mean	98.42	95.77	95.41	96.94	94.65	97.07	97.78	96.06	96.51
	± std	0.04	0.09	0.17	0.09	0.17	0.03	0.05	0.08	0.09
LAN (2019) [32]	mean	98.75	96.26	95.59	97.28	94.94	97.59	**98.04**	96.55	96.88
	± std	0.02	0.12	0.13	0.08	0.11	0.04	0.04	0.01	0.07
Ours	mean	**99.22**	**98.40**	**97.30**	**98.29**	**96.71**	**99.10**	96.80	**98.69**	**98.06**
	± std	0.08	0.02	0.03	0.02	0.05	0.01	0.02	0.04	0.03

TABLE VIII
ABLATION STUDY

Ablation Models	OntoNotes 5.0	CoNLL-2000
Baseline	87.41	94.97
w/o Word Topic Graph	88.27	95.90
w/o Word Co-occurrence Graph	88.45	95.93
w/o Sentence Dependency Graph	88.19	95.84
w/o Graph Aggregation	87.81	95.79
w/o Label-aware Attention Mechanism	88.64	95.94
Full Model	**88.75**	**96.00**

TABLE IX
THE RESULTS OF IV, OOTV, OOEV, OOBV ON THE ONTONOTES 5.0 DATASET

	Baseline			**Our Model**		
	P	R	F_1	P	R	F_1
IV	88.65	87.47	88.06	89.71	89.22	**89.46**
OOTV	86.33	86.81	86.57	88.05	88.43	**88.24**
OOEV	82.92	86.32	84.59	82.83	86.74	**84.73**
OOBV	84.51	87.21	85.84	85.71	87.99	**86.84**

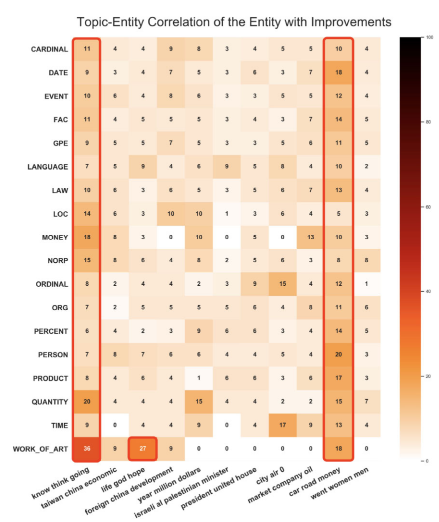

Fig. 4. Percentage of correctly predicted entity types (y axis)(improvements relative to baseline) with respect to corresponding document topics (x axis) in the OntoNotes dataset.

OntoNotes dataset consists of news, magazine and other social media text, so it can be seen that the topics of a large proportion of entities are "know think going" (the first column) and "car road money" (the tenth column). Furthermore, we select the entities correctly predicted by our model but failed in the baseline model. Then we draw its heatmap in Figure 4 to visualize the improvements compared with baseline. Comparing Figure 3 and Figure 4, we can find that they are similar with density. The main improvements come from the topic "know think going" and "car road money". For specific entity types like WORK_OF_ART got improvements from the topic "life god hope" (the third column). Through the comparison, we can conclude that our model truly captures topic-entity relations and gain significant improvements.

TABLE X
THE F1 OF ENTITIES WITH DIFFERENT PMI IN THE ONTONOTES 5.0 DATASET

	Baseline	Our Model	Improvements
Average PMI < 5	89.87	91.31	1.44 ↑
$5 <$ Average PMI < 10	88.74	89.51	0.77 ↑
Average PMI > 10	75.87	80.18	**4.31** ↑

H. Word Co-Occurrence Analysis

To show the relationships between word co-occurrence and sequence labeling, we analysis the words with different co-occurrence information and its results on the OntoNotes dataset in Table X. Firstly, we average the word co-occurrence frequency PMI between the target word and other context words in a fixed window 10 to get its average PMI. Through our analysis, the PMI of 33.1% of the words are < 5, the PMI of 12.2% of the words are > 10. It can be seen that the entities with lower PMI achieves better results. We found that the words with lower PMI value contain lots of PERSON, DATE and TIME entites, which have various forms and fewer co-occurrence information. We speculate that the results of words with lower PMI are higher because it is easier to recognize PERSON, DATE and TIME words. Furthermore, compared with baseline, it can be seen that our model gets the largest improvements from the words with higher co-occurrence information (PMI > 10). We can conclude that word co-occurrence information bring significant improvements especially on the words with rich co-occurrence information.

TABLE XI
SAMPLE RESULT. BASELINE MODEL WRONGLY PREDICTS ENTITY DREAMCAST TO S-ORG WHILE OUR MODEL TRULY PREDICTS IT TO S-PRODUCT

Examples	Last year Sega 's **Dreamcast** blew minds and the competition.
Ground Truth	S-PRODUCT
Baseline	S-ORG
Our Model	S-PRODUCT

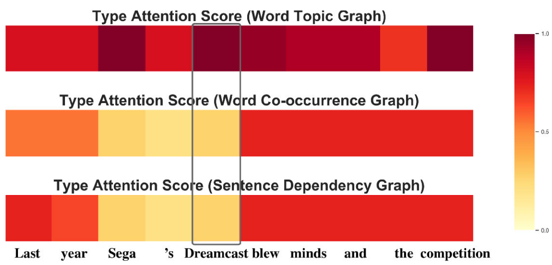

Fig. 5. Attention maps for case instance.

I. Case Analysis

To show how corpus-aware non-sequential features and attention mechanism work, we select a typical case in Table XI and visualize its attention map in Figure 5. In Table XI, baseline method incorrectly recognizes entity *Dreamcast* as type ORGANIZATION, while our model correctly predicts it as type PRODUCT. It is a little confusing for the model to understand the unseen words merely from sequential features. We find the document topic words of the topic with the largest probabilities are "years, product, technology". In Figure 5, phrase *Sega's Dreamcast* has a higher attention score in word topic graph attention map, which demonstrates that our model utilizes topic information to enhance the word representations. For word co-occurrence information, word *Dreamcast* doesn't appear in the training corpus. For sentence dependency graph, the grammar structure *nsubj (nominal subject)* doesn't provide distinguishing information on tagging. So the rest two attention maps demonstrate that our model doesn't pay attention to word co-occurrence information and syntactic dependencies. To summarize, our model models corpus-aware non-sequential features and dynamically aggregates through attention mechanism, which helps us perform tagging with better results.

V. CONCLUSION

Non-sequential features are powerful and effective in sequence labeling tasks. In this work, we propose a corpus-aware graph aggregation network to model non-sequential features both in sentence-level and corpus-level. Word topic graphs, word co-occurrence graph and sentence dependency graph are elaborately designed. We propose a label-aware attention mechanism to fuse different features dynamically. Empirical studies on sequence labeling datasets suggest that our model outperforms previous state-of-the-art systems. In the future, we will explore more available corpus-aware information and its adaptive modeling methods.

REFERENCES

[1] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*.
[2] C. Dong, J. Zhang, C. Zong, M. Hattori, and H. Di, "Character-based LSTM-CRF with radical-level features for chinese named entity recognition," in *Natural Lang. Understanding Intell. Appl.*Berlin, Germany: Springer, 2016, pp. 239–250.
[3] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, (Vol 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1064–1074.
[4] J. Z. J. X. Y. C. Yijin Liu, F. Meng and J. Zhou, "Gcdt: A global context enhanced deep transition architecture for sequence labeling," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2431–2441.
[5] Y. Qian, E. Santus, Z. Jin, J. Guo, and R. Barzilay, "GraphIE: A graph-based framework for information extraction," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, Jun. 2019, pp. 751–761.
[6] Y. Zhang, Q. Liu, and L. Song, "Sentence-State LSTM for Text Representation," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 317–327.
[7] Z. Jie and W. Lu, "Dependency-guided LSTM-CRF for named entity recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 3862–3872.
[8] Y. Luo, F. Xiao, and H. Zhao, "Hierarchical contextualized representation for named entity recognition," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 8441–8448.
[9] O. Adams, A. Makarucha, G. Neubig, S. Bird, and T. Cohn, "Cross-lingual word embeddings for low-resource language modeling," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*: Volume 1, Long Papers. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 937–947. [Online]. Available: https://www.aclweb.org/anthology/E17-1088
[10] V. Krishnan and C. D. Manning, "An effective two-stage model for exploiting non-local dependencies in named entity recognition," in *Proc. 21st Int. Conf. Comput. Linguistics 44th Annu. Meeting Assoc. Comput. Linguistics*. Sydney, Australia: Association for Computational Linguistics, Jul. 2006, pp. 1121–1128. [Online]. Available: https://www.aclweb.org/anthology/P06-1141
[11] S. Pouriyeh, M. Allahyari, K. Kochut, G. Cheng, and H. R. Arabnia, "ES-LDA: Entity summarization using knowledge-based topic modeling," in *Proc. 8th Int. Joint Conf. Natural Lang. Process.* (Volume 1: *Long Papers). Taipei, Taiwan: Asian Federation of Natural Language Processing*, Nov. 2017, pp. 316–325. [Online]. Available: https://www.aclweb.org/anthology/I17-1032
[12] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://www.aclweb.org/anthology/D14-1162
[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
[14] A. Akbik, T. Bergmann, and R. Vollgraf, "Pooled contextualized embeddings for named entity recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, Volume 1 (Long and Short Papers), 2019, pp. 724–728.
[15] J. Luoma and S. Pyysalo, "Exploring cross-sentence contexts for named entity recognition with BERT," in *Proc. 28th Int. Conf. Comput. Linguistics*. Barcelona, Spain (Online): *International Committee on Computational Linguistics*, Dec. 2020, pp. 904–914. [Online]. Available: https://www.aclweb.org/anthology/2020.coling-main.78
[16] Y. Shin and S. Lee, "Learning context using segment-level LSTM for neural sequence labeling," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 105–115, 2020.
[17] M. E. Peters *et al.*, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, Jul. 2018, pp. 2227–2237.
[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North American Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, Jun. 2019, pp. 4171–4186.
[19] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," *Proc. AAAI Conf. Artif. Intell.*, vol. 33. no. 01, pp. 7370–7377, 2019.

[20] D. Marcheggiani and I. Titov, "Encoding sentences with graph convolutional networks for semantic role labeling," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Sep. 2017, pp. 1506–1515.

[21] G. Wang *et al.*, "Joint embedding of words and labels for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics* (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2321–2331.

[22] S. Pradhan *et al.*, "Towards robust linguistic analysis using ontonotes," in *Proc. 17th Conf. Comput. Natural Lang. Learn.*, 2013, pp. 143–152.

[23] E. F. Sang and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," in *Proc. 17th Conf. Natural Lang. Learn. HLT-NAACL*, 2003, pp. 142–147.

[24] E. F. T. K. Sang and S. Buchholz, "Introduction to the CoNLL-2000 shared task chunking," in *Proc. 4th Conf. Comput. Natural Lang. Learn. 2nd Learn. Lang. Log. Workshop*, 2000, pp. 127–132. [Online]. Available: https://www.aclweb.org/anthology/W00-0726

[25] N. Silveira *et al.*, "A gold standard dependency corpus for english," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC'14)*. Reykjavik, Iceland: *European Language Resources Association (ELRA)*, May 2014, pp. 2897–2904. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1089_Paper.pdf

[26] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 task 4: Aspect based sentiment analysis," in *Proc. 8th Int. Workshop Semantic Eval.* (SemEval 2014). Dublin, Ireland: Association for Computational Linguistics, Aug. 2014, pp. 27–35. [Online]. Available: https://www.aclweb.org/anthology/S14-2004

[27] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "SemEval-2015 task 12: Aspect based sentiment analysis," in *Proc. 9th Int. Workshop Semantic Eval.* (SemEval 2015). Denver, Colorado: Association for Computational Linguistics, Jun. 2015, pp. 486–495. [Online]. Available: https://www.aclweb.org/anthology/S15-2082

[28] M. Pontiki *et al.*, "SemEval-2016 task 5: Aspect based sentiment analysis," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 19–30. [Online]. Available: https://www.aclweb.org/anthology/S16-1002

[29] M. Mitchell, J. Aguilar, T. Wilson, and B. Van Durme, "Open domain targeted sentiment," in *Proc. 2013 Conf. Empirical Methods Natural Lang. Process.* Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1643–1654. [Online]. Available: https://www.aclweb.org/anthology/D13-1171

[30] X. Li, L. Bing, P. Li, and W. Lam, "A unified model for opinion target extraction and target sentiment prediction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 6714–6721.

[31] H. Chen, Z. Lin, G. Ding, J. Lou, Y. Zhang, and B. Karlsson, "GRN: Gated relation network to enhance convolutional neural network for named entity recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6236–6243.

[32] L. Cui and Y. Zhang, "Hierarchically-refined label attention network for sequence labeling," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4115–4128. [Online]. Available: https://www.aclweb.org/anthology/D19-1422

[33] X. Wang *et al.*, "Automated concatenation of embeddings for structured prediction," 2020, *arXiv:2010.05006v3*.

[34] X. Li, L. Bing, P. Li, W. Lam, and Z. Yang, "Aspect term extraction with history attention and selective transformation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 4194–4200.

[35] X. Li, L. Bing, W. Lam, and B. Shi, "Transformation networks for target-oriented sentiment classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2018, pp. 946–956.

[36] J. P. Chiu and E. Nichols, "Named entity recognition with bidirectional lstm-cnns," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 357–370, 2016.

[37] Y. Shen, H. Yun, Z. Lipton, Y. Kronrod, and A. Anandkumar, "Deep active learning for named entity recognition," in *Proc. 2nd Workshop Representation Learn. NLP*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 252–256. [Online]. Available: https://www.aclweb.org/anthology/W17-2630

[38] E. Strubell, P. Verga, D. Belanger, and A. McCallum, "Fast and accurate entity recognition with iterated dilated convolutions," in *Proc. 2017 Conf. Empirical Methods Natural Lang. Process.*, EMNLP 2017, Copenhagen, Denmark, Sep. 9-11, 2017, 2017, pp. 2670–2680. [Online]. Available: https://www.aclweb.org/anthology/D17-1283/

[39] A. Ghaddar and P. Langlais, "Robust lexical features for improved neural network named-entity recognition," in *Proc. 27th Int. Conf. Comput. Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1896–1907. [Online]. Available: https://www.aclweb.org/anthology/C18-1161

[40] K. Clark, M.-T. Luong, C. D. Manning, and Q. Le, "Semi-supervised sequence modeling with cross-view training," in *Proc. Conf. Empirical Methods Natural Lang. Process.*. Brussels, Belgium: Association for Computational Linguistics, Oct./Nov. 2018, pp. 1914–1925. [Online]. Available: https://www.aclweb.org/anthology/D18-1217

[41] Y. Nie, Y. Tian, Y. Song, X. Ao, and X. Wan, "Improving named entity recognition with attentive ensemble of syntactic information," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4231–4245. [Online]. Available: https://www.aclweb.org/anthology/2020.findings-emnlp.378

[42] F. Zhai, S. Potdar, B. Xiang, and B. Zhou, "Neural models for sequence chunking," *Proc. 31st AAAI Conf. Artif. Intell.*, vol. 31, no. 1, pp. 3365–3371, 2017.

[43] A. Søgaard and Y. Goldberg, "Deep multi-task learning with low level tasks supervised at lower layers," in *Proc. 54th Annu. Meeting Assoc Comput. Linguistics* (Volume 2: *Short Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 231–235. [Online]. Available: https://www.aclweb.org/anthology/P16-2038

[44] Y. Xin, E. Hart, V. Mahajan, and J.-D. Ruvini, "Learning better internal structure of words for sequence labeling," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2584–2593. [Online]. Available: https://www.aclweb.org/anthology/D18-1279

[45] L. Zhao, X. Qiu, Q. Zhang, and X. Huang, "Sequence labeling with deep gated dual path cnn," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2326–2335, Dec. 2019.

[46] L. Liu *et al.*, "Empower sequence labeling with task-aware neural language model," *Proc. AAAI Conf. Artif. Intell.*. vol. 32. No. 1, pp. 5253–5260, 2018.

[47] M. E. Peters, W. Ammar, C. Bhagavatula, and R. Power, "Semi-supervised sequence tagging with bidirectional language models," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1756–1765.

[48] A. Hu, Z. Dou, J. Nie, and J.-R. Wen, "Leveraging multi-token entities in document-level named entity recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 7961–7968.

[49] X. Li, L. Bing, W. Zhang, and W. Lam, "Exploiting BERT for end-to-end aspect-based sentiment analysis," in *Proc. 5th Workshop Noisy User-generated Text* (W-NUT 2019), 2019, pp. 34–41. [Online]. Available: https://www.aclweb.org/anthology/D19-5505

[50] M. Yasunaga, J. Kasai, and D. Radev, "Robust multilingual part-of-speech tagging via adversarial training," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 976–986. [Online]. Available: https://www.aclweb.org/anthology/N18-1089