Leveraging Multi-Token Entities in Document-Level Named Entity Recognition

Anwen Hu,[2,3] **Zhicheng Dou,**[1,2] **Jian-Yun Nie,**[5] **Ji-Rong Wen**[3,4]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]School of Information, Renmin University of China
[3]Beijing Key Laboratory of Big Data Management and Analysis Methods
[4]Key Laboratory of Data Engineering and Knowledge Engineering, MOE
[5]Department of Computer Science and Operations Research, University of Montreal
{anwenhu, dou, jrwen}@ruc.edu.cn, nie@iro.umontreal.ca

Abstract

Most state-of-the-art named entity recognition systems are designed to process each sentence within a document independently. These systems are easy to confuse entity types when the context information in a sentence is not sufficient enough. To utilize the context information within the whole document, most document-level work let neural networks on their own to learn the relation across sentences, which is not intuitive enough for us humans. In this paper, we divide entities to multi-token entities that contain multiple tokens and single-token entities that are composed of a single token. We propose that the context information of multi-token entities should be more reliable in document-level NER for news articles. We design a fusion attention mechanism which not only learns the semantic relevance between occurrences of the same token, but also focuses more on occurrences belonging to multi-tokens entities. To identify multi-token entities, we design an auxiliary task namely 'Multi-token Entity Classification' and perform this task simultaneously with document-level NER. This auxiliary task is simplified from NER and doesn't require extra annotation. Experimental results on the CoNLL-2003 dataset and OntoNotes$_{\text{nbm}}$ dataset show that our model outperforms state-of-the-art sentence-level and document-level NER methods.

Introduction

Named entity recognition (NER) is one of the first stages for natural language processing. Most neural network based models are designed for sentence-level NER: they treat sentences in a document independently during training or predicting (Huang, Xu, and Yu 2015; Chiu and Nichols 2016; Lample et al. 2016; Ma and Hovy 2016; Gregoric, Bachrach, and Coope 2018; Peters et al. 2018). This may easily lead to tagging inconsistency problems: the same entity in two different sentences might be recognized as different entity types. For the example given in Figure 1, there are three occurrences of the token 'Matsushita' in three sentences within a document. A sentence-level NER model, namely BiLSTM-CNNs-CRF (Ma and Hovy 2016), can successfully recognize 'Yasuo Matsushita' as a 'PERSON'. However, it incorrectly classifies the latter two 'Matsushita' as 'ORG'. After investigation, we find 'Matsushita' usually refers to the organization 'Matsushita Electric Industrial'. In the glove embedding space, it is most close to the names of Japanese companies such as 'sanyo' and 'panasonic'. Thus, without further contextual information, the latter two 'Matsushita' are easily recognized as 'ORG'.

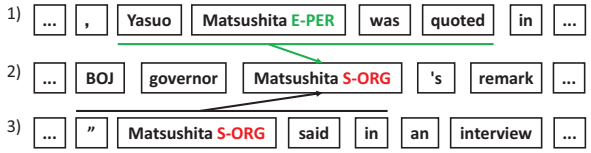


Figure 1: An example of the label inconsistency problem within a document in the CoNLL-2003 English dataset. Green and red tags indicate respectively correct and incorrect tags predicted by a sentence-level model. Green and black arrows refer to useful and less useful contextual information for the second 'Matsushita' token.

A possible solution to solve the above problem is to extend sentence-level NER to document-level NER, which leverages information from the entire document. For the example shown in Figure 1, the first occurrence of the token 'Matsushita' can provide useful information for disambiguating the second occurrence. Various document-level information has been incorporated, ranging from manually designed information (Chieu and Ng 2003; Finkel, Grenager, and Manning 2005; Krishnan and Manning 2006; Kazama and Torisawa 2007) to automatically learned information (Strubell et al. 2017; Luo et al. 2017; Zhang et al. 2018; Xu, Wang, and He 2018). For example, some global attention mechanisms (Zhang et al. 2018; Luo et al. 2017; Xu, Wang, and He 2018) were designed to utilized context information across sentences. In these methods, attention weights are all calculated based on contextual hidden states produced by LSTM. It's not clear enough which context information should be more reliable and it's hard for us humans to understand what the high attention weights are based on in the learned attention mechanisms. In this paper, we call this kind of attention mechanism as semantic attention mechanism.

In this paper, we propose to pay more attention to the token occurrences within multi-token entities - entities composed of multiple tokens, when modeling document level information for entity recognition. Multi-token entities are commonly used in texts. In an article, it is common that a multi-token entity such as 'Yasuo Matsushita' is fully spelled out at the beginning of the article, and then referred to by one of its token (e.g 'Matsushita') later. We think contextual information of this kind of multi-token entities are more helpful for disambiguating other token occurrences because they are usually more specific. As shown in Figure 1, the author uses 'Matsushita's full name when mentioning him for the first time and then uses the last name to refer to him. It is obvious that the contextual information of first 'Matsushita', a part of the multi-token entity 'Yasuo Matsushita', is more important than the third occurrence, which is also ambiguous. After an investigation to the CoNLL-2003 English data set, we found 26.62% of the single-token entities are constituents of multi-token entities in the same document. Furthermore, among these single-token entities, 78.87% of them have at least one multi-token entity of the same type within a document. Multi-token entities are usually less ambiguous than single-token entities. They provide useful information for disambiguating other occurrences of single-token entities in the same document. This is the idea we exploit in this paper.

For news articles, we propose a **ME** (Multi-token Entity)-**I**nformed **D**ocument-level model (**MEID**) for document-level named entity recognition based on the above investigation. Specifically, MEID uses a fusion attention mechanism to generate document-level features. Besides taking into account semantic relevance between occurrence of a particular token, our fusion attention mechanism pays more attention to occurrences which are a part of multi-token entities. However, whether or not a token belongs to a multi-token entity is not informed in the inputs of NER. To introduce this information, we design an auxiliary Multi-token Entity Classification task which is jointly learned with NER and doesn't requires extra annotation. The document-level features are then combined with the local features in sentences, as the input to the CRF layer to decode token labels. We evaluate the model on the CoNLL-2003 dataset and $\text{OntoNotes}_{\text{nbm}}$ dataset. Experiments show our model leveraging multi-token entities can significantly improve the recognition quality.

The main contributions of this paper are:

- We propose a novel attention-based document-level NER model that leverages global context features across sentences as supplements to local context features.
- We take advantage of multi-token entities in the document to guide NER. Multi-token entities are detected by an auxiliary sequence tagging task.
- Experimental results confirm the effectiveness of the proposed method over the state-of-the-art sentence-level and document-level NER models.

Related Work

Sentence-level NER

There are many statistical models successfully applied in sentence-level NER, like HMM (Leek 1997) and CRF (Lafferty, McCallum, and Pereira 2001). In recent years, many neural network based methods (Huang, Xu, and Yu 2015; Lample et al. 2016; Chiu and Nichols 2016; Ma and Hovy 2016) encoded sentences with LSTM for its advantages of modeling sequence data. Besides, some work (Collobert et al. 2011; Yao et al. 2015; Strubell et al. 2017; Wu et al. 2015; Yang, Liang, and Zhang 2018) explored using CNN-based neural networks to encode sentences. Furthermore, many studies (Chiu and Nichols 2016; Lample et al. 2016; Kuru, Can, and Yuret 2016; Gridach 2017; Rei, Crichton, and Pyysalo 2016; dos Santos and Guimarães 2015) proposed to encode character-level features with neural network, such as CNN and LSTM. In this paper, we use the BiLSTM-CNNS-CRF model (Ma and Hovy 2016), a truly end-to-end sequence labeling model, as the basis and focus more on document-level NER.

Document-level NER

Document-level NER models used global information in a document to help entity recognition. Some manually designed non-local features (Chieu and Ng 2003; Finkel, Grenager, and Manning 2005; Krishnan and Manning 2006) were introduced to statistic-based methods and obtained promising results. Context aggregation feature (Ratinov and Roth 2009) was defined as aggregated tokens around occurrences of a particular token, which had higher coverage than manually designed non-local features. Recently, neural networks were employed to automatically learn document-level information. Some work (Strubell et al. 2017; Luo et al. 2017) concatenated sentences within a document and used the CNN based neural network or the attention mechanism to encode the whole long sequence. Attention-based models (Luo et al. 2017; Zhang et al. 2018; Xu, Wang, and He 2018) found supporting document-level context information according to weights calculated from local contextual feature vectors. Attention weights calculated from contextual feature vectors are hard for us humans to understand what they represents. In this paper, our model pays more attention to occurrences belonging multi-token entity, which is more clear and interpretable. The most relevant work to ours is GlobalAtt (Zhang et al. 2018). GlobalAtt applied a BiLSTM to independently encode each sentence. It then utilized a self-attention mechanism among local contextual features of other occurrences of the same token to generate the global feature. There are two main differences between their model and ours. First and foremost, our attention mechanism takes care of occurrences belonging multi-token entities, which can provide useful hints to disambiguate NER in many cases, as we explained earlier. Second, they applied a gate mechanism and a BiLSTM to control the influence of document-level features, and we achieve comparable performance with a single MLP layer.

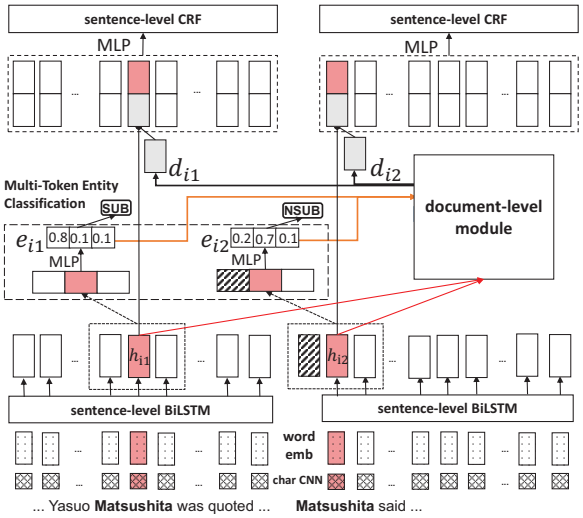


Figure 2: The overall architecture of our model. Sentence-level BiLSTM generates local contextual features $\mathbf{h_i} = (h_{i1}, h_{i2}, ..., h_{ik})$. Multi-token Entity Classification module identifies multi-token entities and outputs corresponding ME features $\mathbf{e_i} = (e_{i1}, e_{i2}, ..., e_{ik})$. Document-level module takes in $(\mathbf{h_i}, \mathbf{e_i})$ and returns document-level features $\mathbf{d_i} = (d_{i1}, d_{i2}, ..., d_{ik})$.

Multi-task Learning

Multi-task learning is widely used in natural language processing, such as jointly learning Chinese word segmentation and named entity recognition (Peng and Dredze 2017), jointly performing aspect detection and sentiment classification (Wang et al. 2019), jointly extracting named entities and relation (Zheng et al. 2017). Semantic role labeling model LISA (Strubell et al. 2018) simultaneously learned dependency parsing and used a syntactically-informed self-attention mechanism to attend to each token's syntactic parse parent. Inspired by this strategy, to find and attend to tokens belonging to multi-token entities within a document, we design an auxiliary task to classify whether a token is a part of a multi-token entity.

Our Document-level Model: MEID

Task Definition

In Document-level NER, each document D is represented as a sequence of sentences $(x_1, x_2, ..., x_s)$, where $x_i (1 \leq i \leq s)$ is represented as a sequence of tokens $(w_1, w_2, ..., w_l)$. For each sentence x, our goal is to use context information of both the sentence and the whole document to predict a tag sequence $y = (y_1, y_2, .., y_l)$, where $y_i (1 \leq i \leq l) \in \mathcal{Y}$. $\mathcal{Y}$ is the set of tags following the BIOES tagging scheme. Except 'O', each tag in $\mathcal{Y}$ is composed of the boundary of the entity and the entity type. For example, 'B-PER' denotes the **b**eginning token of a person name and 'S-LOC' denotes a **s**ingle-token location name.

Model Overview

Our model applies a fusion attention mechanism to focus more on multi-token entities that provide useful information to other ambiguous occurrences. Whether an occurrence belongs to a multi-token entity is not included in the inputs of the NER task and it can only be known after a preliminary recognition. So, in this paper, we introduce a 'Multi-token Entity Classification'(MEC) task without extra annotation to predict whether an occurrence belongs to a multi-token entity. We use multi-task learning to jointly perform MEC and document-level NER.

The overall architecture of our proposed model is shown in Figure 2. In this model, we first encode sentences of a document independently by a sentence-level BiLSTM layer. Then a Multi-token Entity Classification module is used to judge whether a token belongs to a multi-token entity. For a particular token, local contextual features and Multi-token Entity Classification results of its occurrences are fed to a document-level representation learning module. The document-level module returns a document-level feature for each occurrence. We then use a MLP layer to fuse the local contextual features and the document-level features. Finally, we apply a CRF layer at the sentence level to decode the final label sequences. We will describe the details of each component in the remaining parts of this section.

Character-level Representation

Previous studies (Chiu and Nichols 2016; Lample et al. 2016; Yang, Liang, and Zhang 2018) have shown character information such as capitalization can significantly improve sequence labeling models. Besides, experiments in (Yang, Liang, and Zhang 2018) showed that there was no significant difference in using CNN or LSTM to extract character features. Therefore, in this paper, we use character-level CNN (Ma and Hovy 2016) to extract morphological information of a given word.

Sentence-level Context Representation

In this paper, we apply the Long Short-Term Memory Network (Hochreiter and Schmidhuber 1997) (LSTM) to encode the token sequences within each sentence. LSTM is developed from recurrent neural network (RNN) and is better to handle long-term dependencies. To take into account context information before and after, BiLSTM concatenates two LSTMs in forward direction $\overrightarrow{h_t}$ and in backward direction $\overleftarrow{h_t}$. The state of each token is represented by:

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]. \quad (1)$$

In this paper, we call h_t produced by sentence-level BiLSTM as local/sentence-level contextual representation. For each token w_i, we record location information of its all occurrences within a document as :

$$\mathbf{u_i} = \{(s_{i1}, o_{i1}), (s_{i2}, o_{i2}), ..., (s_{ik}, o_{ik})\},$$

where k is the count of occurrences for token w_i, (s_{it}, o_{it}) means the t^{th} occurrence is the ${o_{it}}^{th}$ token in the ${s_{it}}^{th}$ sentence of a document. Then, according to these location infor-

mation of all occurrences for w_i, we obtain a list of sentence-level BiLSTM outputs for $\mathbf{u_i}$:

$$\mathbf{h_i} = (h_{i1}, h_{i2}, ..., h_{ik}), \tag{2}$$

where $h_{it}(1 \leq t \leq k)$ is obtained by Eq. (1), which is the sentence-level contextual representation of the t^{th} occurrence of token w_i; $\mathbf{h_i}$ is one of the inputs for the document-level module for token w_i.

Multi-token Entity Classification

In this paper, we use Multi-token Entity (ME) information to refer to whether an occurrence belongs to a multi-token entity. To inform the model accurate ME information of each occurrence, we add an auxiliary supervised task called Multi-token Entity Classification(MEC) based on the original NER task. In BIOES tagging scheme used in the NER task, 'B-','I-', 'E-' refer to the 'Begin', 'Inside' and 'End' of an entity, 'S-' refers to a single-token entity, 'O' refers to non-entity token. So, if a token is labeled as 'B-','I-' or 'E-', it belongs to a multi-token entity and we label it as 'SUB' (**sub** token of an entity) in ME classification task. If a token is labeled as 'S-', it is a single-token entity and we label it as 'NSUB'. 'O' tags are retained from the NER task. Thus, the ME classification task is a simple three-class classification task. We share the parameters of sentence-level BiLSTM layers in our model for the MEC task and NER task. For the MEC task, we concatenate the previous and the next sentence-level contextual representation with the target sentence-level contextual representation as $\overline{h}_t$. We set the window size to 3 because 98.33% and 85.31% of entities have fewer than 4 tokens in CoNLL-2003 dataset and OntoNotes$_{\text{nbm}}$ dataset respectively. Then we feed $\overline{h}_t$ to MLP to predict ME tags:

$$\overline{h}_t = [h_{t-1}, h_t, h_{t+1}], \tag{3}$$

$$e_t = \text{Softmax}(\text{MLP}(\overline{h}_t)), \tag{4}$$

where e_t is the predicted probability distribution after the softmax layer. Given a sentence $\mathbf{x} = (w_1, w_2, ..., w_l)$, the cross-entropy loss for MEC task is defined as:

$$\text{loss}_M = -\sum_{t=1}^{l} e'_t \log(e_t), \tag{5}$$

where e'_t (one-hot vector) is the true distribution for w_t.

To assist document-level NER, we obtain a list of ME representations for $\mathbf{u_i}$:

$$\mathbf{e_i} = (e_{i1}, e_{i2}, ..., e_{ik}), \tag{6}$$

where $e_{it}(1 \leq t \leq k)$ is obtained by Eq. (4). $\mathbf{e_i}$ is sent to the document-level module, together with $\mathbf{h_i}$ defined in Eq. (2).

Document-level Representation Learning

After getting local contextual representations and ME representation of occurrences for each token, we apply a document-level module to learn the relation between these occurrences and yield the final document-level representation. The architecture of the module is shown in Figure 3.

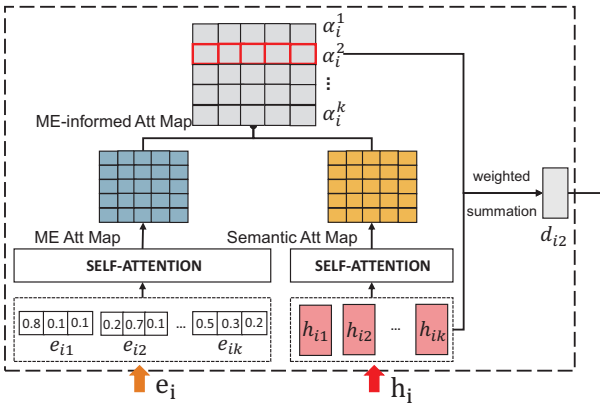


Figure 3: ME-informed Attention in the document-level module. α_i^2 is the distribution of ME-informed attention weight for u_{i2} (the 2^{nd} occurrence of token w_i). d_{i2} is the document-level representation for u_{i2}.

ME-informed Attention It is difficult for neural network to learn the complex relation between occurrences of a particular token on its own. Besides, the distributions of attention weights completely computed by semantic features are hard for us humans to understand. In this paper, we propose to focus more on occurrences belonging to multi-token entities based on our investigation. To achieve this target, we add a weak guide when generating document-level features by a fusion attention mechanism called ME-informed attention. Besides semantic self-attention, we apply another attention mechanism on ME representations $\mathbf{e_i}$ to attend more to occurrences belonging to multi-token entities. The semantic attention score and ME attention score are calculated as the attention mechanism (Bahdanau, Cho, and Bengio 2015):

$$s_{in}^m = v_s^\top \tanh(W_{s_1} h_{im} + W_{s_2} h_{in} + b_s)(h_{im}, h_{in} \in \mathbf{h_i}),$$

$$u_{in}^m = v_u^\top \tanh(W_{u_1} e_{im} + W_{u_2} e_{in} + b_u)(e_{im}, e_{in} \in \mathbf{e_i}),$$

where $\mathbf{h_i}$ and $\mathbf{e_i}$ is obtained by Eq. (2) and Eq. (6); s_{in}^m and u_{in}^m are respectively the semantic attention score and ME attention score of the n^{th} occurrence for the m^{th} occurrence of token w_i. Then we fuse these two attention scores and weight local contextual features for each occurrence:

$$\alpha_{in}^m = \text{Softmax}(s_{in}^m + u_{in}^m),$$

$$\alpha_{\mathbf{i}}^{\mathbf{m}} = (\alpha_{i1}^m, \alpha_{i2}^m, ..., \alpha_{ik}^m),$$

$$d_{im} = \sum_{n=1}^{k} \alpha_{in}^m h_{in}, (\alpha_{in}^m \in \alpha_{\mathbf{i}}^{\mathbf{m}}, h_{in} \in \mathbf{h_i})$$

where α_{in}^m is the ME-informed attention weight. d_{im} is the document-level representation for the m^{th} occurrence.

Tag Prediction

We apply the Condition Random Fields (CRF) (Lafferty, McCallum, and Pereira 2001) for the tag prediction step. CRF has proven to be very effective in sequence labeling tasks, because it considered the correlations between

labels in neighborhoods and decoded the best label sequence rather than a single best label. For a sentence $\mathbf{x} = (w_1, w_2, ..., w_l)$, after concatenating sentence-level contextual representations and document-level representations, we get new hidden states $\hat{\mathbf{h}_\mathbf{x}} = (\hat{h_1}, \hat{h_2}, ..., \hat{h_l})$. Before the CRF layer, we apply an MLP layer to reduce the dimension of vectors in $\hat{\mathbf{h}_\mathbf{x}}$ to e, where e is the number of distinct tags.

For any possible predicted sequence $\mathbf{y} = (y_1, y_2, .., y_l)$, its score is defined as:

$$s(\mathbf{x}, \mathbf{y}) = \sum_{i=0}^{l} A_{y_i, y_{i+1}} + \sum_{i=1}^{l} P_{i, y_i},$$

where $\mathbf{A}$ is a transition matrix, whose element $A_{i,j}$ represents the transition score from tag i to tag j; P_{i,y_i} refers to the score of the y_i tag of the i^{th} token in the sentence. Then a softmax is used to compute the probability of sequence $\mathbf{y}$:

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp s(\mathbf{x}, \mathbf{y})}{\sum_{\tilde{\mathbf{y}} \in \mathbf{Y_x}} \exp s(\mathbf{x}, \tilde{\mathbf{y}})},$$

where $\mathbf{Y_x}$ is a set of all possible tag sequences. During training, we maximize the log-probability of the true tag sequence. The loss function for the NER task is defined as:

$$\text{loss}_N = -\log(p(\hat{\mathbf{y}}|\mathbf{x})),$$

where $\hat{\mathbf{y}}$ is the true tag sequence. Given loss_N and loss_M (the ME classification loss defined in Eq. (5)), the overall loss of MEID, which is a joint loss of the NER and the ME classification task, is defined as:

$$\text{loss} = \text{loss}_M + \text{loss}_N.$$

While decoding, we predict the NER tag sequence which obtains the maximum score:

$$\mathbf{y}* = \arg\max_{\tilde{\mathbf{y}} \in \mathbf{Y_x}} s(\mathbf{x}, \tilde{\mathbf{y}})$$

Experiments

Dataset and Settings

We conduct experiments on the CoNLL-2003 and $\text{OntoNotes}_{\text{nbm}}$ dataset. CoNLL-2003 dataset (Sang and Meulder 2003) is part of the Reuters Corpus comprised of news stories between August 1996 and August 1997. It contains named entities of 4 different types: persons (PER), organizations (ORG), locations (LOC), and miscellaneous names (MISC). We build $\text{OntoNotes}_{\text{nbm}}$ dataset by combining **n**ewswire (nw), **b**roadcast news (bn) and **m**agazine (mz) parts of OntoNotes 5.0 (Hovy et al. 2006). Documents in telephone conversation (TC), web data (WB) and pivot text (PT) parts of OntoNotes 5.0 may not contain global relevance like news stories. Thus we think these three parts are inappropriate for document-level NER and drop them in our experiments. $\text{OntoNotes}_{\text{nbm}}$ dataset contains 18 named entity types. As for tagging scheme, we choose the BIOES, which has been proven better than BIO2 by previous studies (Ratinov and Roth 2009; Lample et al. 2016). The statistics of the two datasets are shown in Table 1.

In this paper, we use the standard F1-score (F1) as the evaluation metrics. Due to the small size of CoNLL-2003 dataset, we conduct each experiment 5 times on it with different random seeds and report its mean. For $\text{OntoNotes}_{\text{nbm}}$, we run each model once. We use traditional GloVe embedding (Pennington, Socher, and Manning 2014) as our default word embedding. To compare with state-of-the-art pre-trained language models, we also experiment with bert-base (Devlin et al. 2019) and flair (Akbik, Blythe, and Vollgraf 2018) as initialized word embedding.

Baselines

We compare our model with several state-of-the-art NER models using no external knowledge, such as lexicons. These models can be categorized into sentence-level models and document-level models. Besides, we experiment with some variants of MEID to show the effectiveness of our fusion attention mechanism.

Sentence-level models

- **LSTM+CRF** (Lample et al. 2016), which extracts the character-level features with a BiLSTM layer and encodes sentences with another BiLSTM layer.
- **BiLSTM+CNNS+CRF** (Ma and Hovy 2016), which extracts morphological features from characters of word with a CNN and encodes sentences with a BiLSTM layer.
- **ParallelRNNs** (Gregoric, Bachrach, and Coope 2018), which splits a single LSTM to multiple equal-sized smaller ones to reduce parameters and concatenates their outputs before decoding.
- **HSCRFs(JNT)** (Ye and Ling 2018), which designs a Hybrid semi-Markov CRF and jointly decodes labels using CRFs and HSCRFs.

Document-level models

- **Att+BiLSTM+CRF** (Luo et al. 2017), which applies a global attention mechanism among all tokens within a document to extract the document-level feature.
- **IDCNN** (Strubell et al. 2017), which encodes the long sequence concatenated by sentences of a document by iteratively applying a stack of dilated convolutions.
- **GlobalAtt** (Zhang et al. 2018), which calculates semantic attention weights among occurrences of an identical token and uses a gate and a BiLSTM to control the influence of document-level features.

Other baselines

- **SENT**, which is the basic sentence-level model in MEID. It has the same architecture as BiLSTM-CNNS-CRF. The difference between SENT and BiLSTM-CNNS-CRF is the training strategy. The former feeds an entire document as a batch like MEID, and the latter feeds several sentences that may come from different documents.
- **MEID-ME**, which drops the ME part in our model. The one difference between MEID-ME and GlobalAtt is that the former only uses a MLP to fuse sentence-level and document-level features, whereas the latter uses a gate mechanism and a BiLSTM layer.

Dataset		#doc	#sent	#token
CoNLL-2003	Train	946	14,987	20,3621
	Dev	216	3,466	51,362
	Test	231	3,684	46,435
$\text{Ontonotes}_{\text{nbm}}$	Train	3,276	54,971	1,249,399
	Dev	469	8,187	188,654
	Test	284	4,718	105,056

Table 1: Statistics of CoNLL-2003 and $\text{OntoNotes}_{\text{nbm}}$ dataset. **#doc**, **#sent** and **#token** refer to counts of document, sentence and token respectively.

Model	CoNLL-2003	$\text{OntoNotes}_{\text{nbm}}$
LSTM+CRF	90.94	87.57
BiLSTM+CNNS+CRF	91.21	88.42
ParallelRNNs	91.48	85.54
HSCRFs(JNT)	91.38	87.74
Att+BiLSTM+CRF	90.49	88.88
IDCNN	90.65	85.24
GlobalAtt	91.43	88.78
SENT	90.92	88.64
MEID-SEM	91.71	88.71
MEID-ME	91.78	88.84
MEID	**91.92**	**89.16**

Table 2: F1-scores of different models with traditional word embedding (e.g. glove) on CoNLL-2003 and $\text{OntoNotes}_{\text{nbm}}$.

Dataset	Model	glove	bert-base	flair
CoNLL-2003	SENT	90.92	90.66	92.59
	MEID	**91.92**	**91.47**	**93.09**
$\text{OntoNotes}_{\text{nbm}}$	SENT	88.64	88.41	89.89
	MEID	**89.16**	**88.96**	**90.29**

Table 3: F1 scores of our approaches using different word embedding on CoNLL-2003 and $\text{OntoNotes}_{\text{nbm}}$. **glove** refers to conventional glove embedding (Pennington, Socher, and Manning 2014). **flair** means stacking embedding concatenated by glove embedding and contextual string embedding (Akbik, Blythe, and Vollgraf 2018). **bert-base** refers to embeddings produced by bert-base model (Devlin et al. 2019).

- **MEID-SEM**, which drops the semantic attention in our document-level module and only uses ME attention to weight local contextual features.

Overall Results

To verify the effectiveness of MEID, we compare our model with state-of-the-art sentence-level and document-level NER models. Experimental results are shown in Table 2 and Table 3. We find:

(1) As shown in Table 2, **MEID outperforms existing sentence-level models and document-level models at the condition of using conventional word representations like Glove embedding**. SENT achieves comparable performance with BiLSTM-CNNS-CRF. It proves that for the sentence-level model, there is no obvious difference between shuffling all sentences and shuffling documents during training. Thus, the improvement in F1 has nothing to do with our training strategies. Adding our document-level representation on the basis of the sentence-level model can indeed improve the NER quality.

(2) **MEID outperforms MEID-ME and GlobalAtt on both two datasets**. MEID-ME achieves comparable performance with GlobalAtt. It indicates that using a gate mechanism and a BiLSTM to control the influence of document-level information may be redundant. After introducing the ME part, MEID performs better than MEID-ME. It proves that the ME-informed attention mechanism indeed helps find more reliable contextual information across sentences.

(3) **MEID outperforms MEID-SEM**. This shows that only ME attention may not be able to fit complex relationships between occurrences across sentences very well. Combining the semantic attention and the ME attention can achieve better performance than using either of them alone.

(4) As shown in Table 3, **whether using bert-base or flair as initialized word embedding, MEID outperforms SENT**. This indicates that our document-level representation can further increase the F1 score at the base of word representations produced by state-of-the-art pre-trained language models.

To explore whether the improvement given by ME is uniform for all entity types, we compared f1 scores of each entity type between MEID and MEID-ME on the both two datasets. It turns out that on the CoNLL-2003, MEID achieves better results than MEID-ME on 100%(4/4) entity types. On the $\text{Ontonotes}_{\text{nbm}}$, MEID achieves better results than MEID-ME on 72%(13/18) entity types. Therefore, we think the improvements given by contexts of multi-token entities are suitable for most of the entity types.

Besides, to see whether our model will work on other types of documents besides news articles, we build the $\text{Ontonotes}_{\text{c}}$ dataset by combining broadcast conversation (bc) and telephone conversation (tc) parts of OntoNotes5.0. It turns out that MEID-ME (scores 75.17) achieve better results than MEID (scores 74.89). Thus, our proposal doesn't seem to work well on the conversation dataset where documents may not contain global relevance like news articles.

About the computation cost, compared with the sentence-level model BiLSTM-CNNS-CRF, the average time for processing the whole CoNLL-2003 test data increases from 0.496 minutes to 0.523 minutes. The average time for processing $\text{OntoNotes}_{\text{nbm}}$ test data increases from 0.979 minutes to 1.173 minutes. We think this is acceptable.

ME Classification

For the MEC task we described in Section **Multi-token Entity Classification**, our end-to-end MEID model achieves 96.20 F1 score and 92.84 F1 score on the CoNLL-2003 and $\text{OntoNotes}_{\text{nbm}}$ test dataset respectively. We also experiment with the separate MEC task without the multi-task learning with the NER. It gets 95.89 F1 score and 91.99 F1 score

respectively, either of which is lower than the corresponding F1 score of multi-task learning. This shows that the joint learning can help improve ME classification.

Case Analysis

To show how ME works, we select two typical cases from the test set of CoNLL-2003. Figure 4 compares the tags given by MEID-ME and MEID. Figure 5 shows semantic attention maps given by MEID-ME and ME-informed attention maps given by MEID.

In the first example, all occurrences of 'Matsushita' refer to a person's name 'Yasuo Matsushita' and only the first mention of 'Matsushita' is part of a multi-token entity 'Yasuo Matsushita'. The second occurrence of 'Matsushita' is wrongly predicted as 'S-ORG' by MEID-ME. Figure 5a shows MEID-ME already pays attention to the first 'Matsushita' to collect document-level contextual information for the second 'Matsushita', but the attention weight doesn't seem to be big enough to trust the first 'Matsushita'. After informed ME features, MEID pays much more attention to the first 'Matsushita' and correctly classifies the second occurrence as 'S-PER', as shown in Figure 5b.

In the second example, the first and second 'zimbabwe' belong to the multi-token entity 'zimbabwe open', which is an open championship. The local contextual information of the first 'zimbabwe' is weak because it appears in a short title. Figure 5c shows that when collecting document-level contextual information for the first 'ZIMBABWE', MEID-ME pays more attention to the last three 'Zimbabwe', which refer to location names. Therefore, MEID-ME wrongly predicts the first 'ZIMBABWE' as 'B-LOC'. After informed ME features, MEID pays much more attention to the second 'Zimbabwe' and corrects 'B-LOC' to 'B-MISC'. Besides, for the last three 'Zimbabwe', MEID pays much attention to the first and second occurrence, but document-level features do not mislead the final prediction. We think this is because their local contextual features play a bigger role than the global features in this case. This is why we need to leverage both the global and the local features, but not just rely on one of them.

In these two examples, it's hard for us to understand the semantic attention map. But the distributions of ME-informed attention weight meet our expectations and are more interpretable.

Conclusion

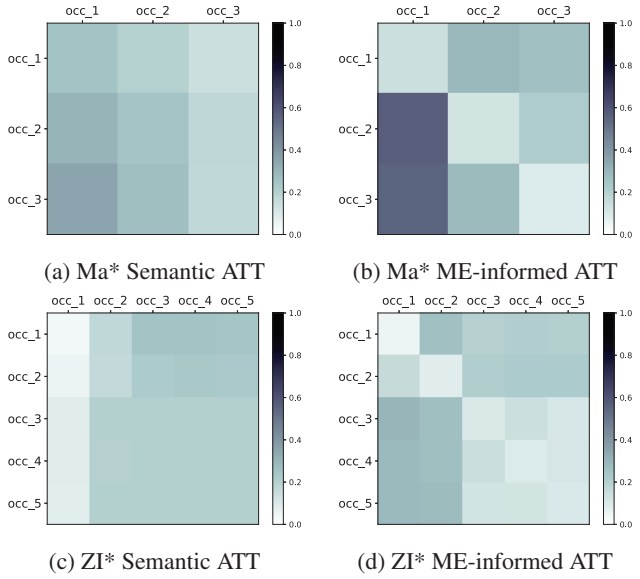
In this paper, we propose a novel neural network designed for document-level NER that takes into account whether occurrences belong to multi-token entities. It doesn't rely on any manually designed non-local features or need extra annotation. We first apply a BiLSTM layer to encode sentences within a document independently. Then we introduce a MLP based module to judge whether a token belongs to a multi-token entity. Meanwhile, we design a document-level module to generate document-level representations from local contextual representations of all occurrences of a particular token. In the document-level module, we apply a ME-informed attention to attend more to multi-token entities.

Sentences in a Document

1) In a rare expression of a view on currencies by the Bank of Japan (BOJ) governor , Yasuo **Matsushita** *E-PER* was quoted in Japan 's leading economic daily on Friday as ...
2) ... was BOJ governor **Matsushita** *S-ORG S-PER* 's remark .
3) ..." **Matsushita** *S-PER* said in an interview with the ...

1) GLF – **ZIMBABWE** *B-LOC B-MISC* OPEN SECND RUND SCRES .
2) Leading second round scores in the **Zimbabwe** *B-MISC* Open at the par-72 Chapman Golf Club on Friday : 132 Des Terblanche 65 67 133 Mark McNulty (**Zimbabwe** *S-LOC*) 72 61 134 Steve van Vuuren 65 69 136 Nick Price (**Zimbabwe** *S-LOC*) 68 68 , Justin Hobday 71 61
3) Andrew Pitts (U.S.) 69 67 138 Mark Cayeux (**Zimbabwe** *S-LOC*) 69 69 , ...

Figure 4: Sample NER results. Blue tags mean correct tags predicted by both MEID-ME and MEID. Red tags mean wrong tags predicted by MEID-ME. Green tags mean correct tags predicted by MEID.



(a) Ma* Semantic ATT (b) Ma* ME-informed ATT

(c) ZI* Semantic ATT (d) ZI* ME-informed ATT

Figure 5: Attention Maps for 'Matsushita'(Ma*) and 'ZIMBABWE'(ZI*). occ_i means the i^{th} occurrence.

Our model MEID achieves state-of-the-art results on the CoNLL-2003 English dataset and OntoNotes$_{\text{nbm}}$ dataset.

Acknowledgments

The corresponding author of this work is Zhicheng Dou. This work was supported by National Natural Science Foundation of China No. 61872370, Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, and the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China No. 2112018391.

References

Akbik, A.; Blythe, D.; and Vollgraf, R. 2018. Contextual string embeddings for sequence labeling. In *COLING*, 1638–1649. Association for Computational Linguistics.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Chieu, H. L., and Ng, H. T. 2003. Named entity recognition with a maximum entropy approach. In *CoNLL*, 160–163. ACL.

Chiu, J. P. C., and Nichols, E. 2016. Named entity recognition with bidirectional lstm-cnns. *TACL* 4:357–370.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. P. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* 12:2493–2537.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 4171–4186. Association for Computational Linguistics.

dos Santos, C. N., and Guimarães, V. 2015. Boosting named entity recognition with neural character embeddings. In *NEWS@ACL*, 25–33. Association for Computational Linguistics.

Finkel, J. R.; Grenager, T.; and Manning, C. D. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, 363–370. The Association for Computer Linguistics.

Gregoric, A. Z.; Bachrach, Y.; and Coope, S. 2018. Named entity recognition with parallel recurrent neural networks. In *ACL (2)*, 69–74. Association for Computational Linguistics.

Gridach, M. 2017. Character-level neural network for biomedical named entity recognition. *Journal of biomedical informatics* 70:85–91.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Hovy, E. H.; Marcus, M. P.; Palmer, M.; Ramshaw, L. A.; and Weischedel, R. M. 2006. Ontonotes: The 90% solution. In *HLT-NAACL*. The Association for Computational Linguistics.

Huang, Z.; Xu, W.; and Yu, K. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR* abs/1508.01991.

Kazama, J., and Torisawa, K. 2007. A new perceptron algorithm for sequence labeling with non-local features. In *EMNLP-CoNLL*, 315–324. ACL.

Krishnan, V., and Manning, C. D. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *ACL*. The Association for Computer Linguistics.

Kuru, O.; Can, O. A.; and Yuret, D. 2016. Charner: Character-level named entity recognition. In *COLING*, 911–921. ACL.

Lafferty, J. D.; McCallum, A.; and Pereira, F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 282–289. Morgan Kaufmann.

Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 2016. Neural architectures for named entity recognition. In *HLT-NAACL*, 260–270. The Association for Computational Linguistics.

Leek, T. R. 1997. Information extraction using hidden markov models. Master's thesis, University of California, San Diego.

Luo, L.; Yang, Z.; Yang, P.; Zhang, Y.; Wang, L.; Lin, H.; and Wang, J. 2017. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics* 34(8):1381–1388.

Ma, X., and Hovy, E. H. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *ACL (1)*. The Association for Computer Linguistics.

Peng, N., and Dredze, M. 2017. Multi-task domain adaptation for sequence tagging. In *Rep4NLP@ACL*, 91–100. Association for Computational Linguistics.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, 1532–1543.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *NAACL-HLT*, 2227–2237. Association for Computational Linguistics.

Ratinov, L., and Roth, D. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL*, 147–155. ACL.

Rei, M.; Crichton, G. K. O.; and Pyysalo, S. 2016. Attending to characters in neural sequence labeling models. In *COLING*, 309–318. ACL.

Sang, E. F. T. K., and Meulder, F. D. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *CoNLL*, 142–147. ACL.

Strubell, E.; Verga, P.; Belanger, D.; and McCallum, A. 2017. Fast and accurate entity recognition with iterated dilated convolutions. In *EMNLP*, 2670–2680. Association for Computational Linguistics.

Strubell, E.; Verga, P.; Andor, D.; Weiss, D.; and McCallum, A. 2018. Linguistically-informed self-attention for semantic role labeling. In *EMNLP*, 5027–5038. Association for Computational Linguistics.

Wang, Y.; Sun, A.; Huang, M.; and Zhu, X. 2019. Aspect-level sentiment analysis using as-capsules. In *The World Wide Web Conference*, 2033–2044. ACM.

Wu, Y.; Jiang, M.; Lei, J.; and Xu, H. 2015. Named entity recognition in chinese clinical text using deep neural network. *Studies in health technology and informatics* 216:624.

Xu, G.; Wang, C.; and He, X. 2018. Improving clinical named entity recognition with global neural attention. In *APWeb/WAIM (2)*, volume 10988 of *Lecture Notes in Computer Science*, 264–279. Springer.

Yang, J.; Liang, S.; and Zhang, Y. 2018. Design challenges and misconceptions in neural sequence labeling. In *COLING*, 3879–3889. Association for Computational Linguistics.

Yao, L.; Liu, H.; Liu, Y.; Li, X.; and Anwar, M. W. 2015. Biomedical named entity recognition based on deep neutral network. *Int. J. Hybrid Inf. Technol* 8(8):279–288.

Ye, Z., and Ling, Z. 2018. Hybrid semi-markov CRF for neural sequence labeling. In *ACL (2)*, 235–240. Association for Computational Linguistics.

Zhang, B.; Whitehead, S.; Huang, L.; and Ji, H. 2018. Global attention for name tagging. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 86–96.

Zheng, S.; Wang, F.; Bao, H.; Hao, Y.; Zhou, P.; and Xu, B. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *ACL (1)*, 1227–1236. Association for Computational Linguistics.