Context-aware Feature Attention Model for Coreference Resolution on Biomedical Texts

Yufei Li\textsuperscript{1,2,3}, Xiangyu Zhou\textsuperscript{1,2,3}, Jie Ma\textsuperscript{1,2,3}, Xiaoyong Ma\textsuperscript{1,2,3} and Chen Li\textsuperscript{1,2,3*}

Abstract

**Background**: Bio-entity Coreference resolution is an important task to gain a complete understanding of biomedical texts automatically. Previous neural network-based studies on this topic are domain system based methods which rely on some domain-specific information integration. However, for the identical mentions, this may lead to misleading information, as the model tends to get similar or even identical representations, which further leads to wrongful predictions.

**Results**: we propose a new context-aware Feature Attention model to distinguish identical mentions effectively to better resolve coreference. The new model can represent identical mentions based on different contexts by adaptively exploiting features effectively. The proposed model substantially outperforms the state-of-the-art baselines on the BioNLP dataset with a 64.0\% F1 score and further demonstrates superior performance on the differential representation and coreferential link of identical mentions.

**Conclusion**: The context-aware Feature Attention model adaptively exploit features and represent identical mentions according to different contexts, which significantly makes the system obtain semantic information effectively and make more accurate predictions. Considering that this approach is still limited when context information is insufficient, we expect to utilize such information more fine-grained with the help of the external knowledge base in coreference resolution.

**Keywords**: Context Awareness; Semantics; Biomedical Text Mining; Coreference Resolution

1School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, P.R. China

Correspondence: cli@xjtu.edu.cn

Full list of author information is available at the end of the article
more than half. Moreover, among them, Wh-qualifiers, prepositions, conjunctions, and noun phrase take up more. In Table 2, we count the identical mentions in each document. We display the percentage of the documents with or without identical mentions and the percentage of documents containing different numbers of the identical mentions. The results show that documents containing identical mentions account for about half of the total documents, of which two identical mentions account for the largest proportion.

Inspired by the recent success of attention mechanism \cite{11, 12}, this paper explores how to distinguish similar or identical text units effectively based on contexts and develops a general context-aware Feature Attention mechanism to enhance the neural biomedical coreference system. Based on a general neural coreference system \cite{13}, we first run the model with integrating some domain-specific information (embeddings pre-trained on large scale corpus, some biomedical and grammatical features). Then we use Feature Attention mechanism to learn the importance or weight of each feature based on contexts such that the span representation can be improved. In this way, the model can adaptively exploit features and differentiate identical mentions based on different contexts.

The proposed model is evaluated on the BioNLP Protein Coreference dataset \cite{14}. The results show that our work achieves state-of-the-art results with a 64.0% F1 score (+2.0 F1). Besides, experimental analysis shows the Feature Attention mechanism does help in distinguishing the identical mentions based on different contexts, which further helps in coreferential links. Moreover, additional experiments on the coreferential linking performance of identical mentions further illustrate the effectiveness of our model.

Related Work

In bio-entity coreference resolution tasks, words referring to each other are called mentions, while a mention can either be a common noun, a proper noun, or a pronoun. Taking the above example in Figure 1, a coreference system partitions the mentions into two coreference chains: ("Protein phosphorylation", "it", "it"), and ("TPCK", "it"). In recent years, several supervised approaches have been proposed for bio-entity coreference resolution. The work can be categorized into three classes. 1) rule and feature-based models \cite{4, 5, 6}, which heavily rely on syntactic parsers to extract manually crafted features and rules. 2) hybrid models \cite{7, 8}, which combine rule-based and machine learning-based methods for biomedical coreference resolution 3) neural network-based models \cite{6, 9, 10}, which use deep learning or neural networks to solve the problem automatically with domain-specific information integration, typically by pre-trained embeddings and some biomedical features. Generally, the above work is summarized in Table 3. Our work is most closely related to the work of \cite{9}, while we focus on the problem that identical mentions tend to get similar or even identical representations, which further mislead to make coreferential mistakes.

Methods

Task

In an end-to-end coreference resolution system, the input is a document \( D \) with \( T \) words, and the output is a set of mention clusters. Let \( N \) be the number of possible text spans in \( D \). We consider all possible spans up to a predefined maximum width. \( \text{START}(i) \) and \( \text{END}(i) \) are the start and end indices of a span \( i \) in \( D \) respectively. For each span \( i \) the system needs to assign an antecedent \( a_i \in \{\epsilon, 1, \ldots, i-1\} \) from all preceding spans or a dummy antecedent \( \epsilon \). The dummy antecedent represents two cases: (1) the span \( i \) is not a mention, or (2) the span \( i \) is a mention but not coreferential with any previous span. Finally, all spans that are connected by a set of antecedent predictions are grouped.

The formal definition of identical mention is as follows. Suppose the \( N \) mentions in a document \( D \) are denoted as \( M = \{m_1, m_2, \ldots, m_N\} \). The identical mentions are defined by: \( M' = \{m_i|\exists \ m_j = m_i \text{ and } m_j \in M \text{ and } j \neq i\} \). For each identical mention \( m_i \), we define its frequency as the number of times that this mention appears in the document.

Baseline Model

In this section, we briefly describe the baseline model: \cite{13} which we will later augment with Feature Attention mechanism.

\textbf{Span Representation}

Assuming vector representation of a sentence with \( L \) words as \( \{x_1, x_2, \ldots, x_L\} \), while \( x_t \) denotes the concatenation of fixed pre-trained word embeddings and CNN character embeddings \cite{15} for \( t-th \) word. The Bidirectional LSTMs \cite{16} are used to encode each \( x_t \).

Then, the model uses the attention mechanism \cite{17} over words in each span to learn a task-specific notion of headedness, and the final representation \( g_i \) of span \( i \) is produced by:

\[
g_i = [x^*_{\text{START}(i)}, x_{\text{END}(i)}^*, \hat{x}_i, \varphi(i)]
\]

where \( x^* \) is the output of Bi-LSTM and \( \hat{x}_i \) is the head embedding encoded by the head attention mechanism. \( \varphi(i) \) is the feature vector encoding the width of the span.
Scoring
The scoring functions: mention score $s_m$ and antecedent score $s_a$ take the span representations as input:

$$s_m(i) = w_m \cdot FFNN_m(g_i)$$
$$s_a(i,j) = w_a \cdot FFNN_a([g_i, g_j, g_i \odot g_j, \varphi^*(i,j)])$$

where $w_m$ and $w_a$ are the weight matrix, $\odot$ denotes element-wise multiplication, FFNN is the feed-forward neural network, and $\varphi(i,j)$ is the pair-wise features encoding the distance between the two spans.

Feature Attention
Domain-specific Information Integration
Similar as [9], to integrate domain-specific information, we make the following variations: 1) For word embedding and head word embedding, we use BioBERT [18] representations pre-trained by the language model BERT [19] as the input to the LSTMs. 2) We integrate one more feature, mentions’ grammatical numbers, to span representation, i.e., whether each mention is singular, plural, or unknown. 3) According to the UMLS semantic types, we employ MetaMapLite [2] to identify the entity’s Metamap type and assign it to all tokens in the entity and encode it as the span’s MetaMap feature.

Model Structure
To avoid the similar representations of identical mentions, we propose a context-aware attention mechanism called Feature Attention to adaptively exploit features based on context.

As shown in Figure 2, we use a general attention mechanism that learns the importance or weight of each feature based on contexts. Suppose the initial feature vectors is $\varphi = [\varphi_1, \varphi_2, \ldots, \varphi_V]$, where $\varphi_j$ indicates the $j$-th feature and $x_u$ is the contexts vectors generated by Bi-LSTM for span $u$ (here we use the average of the context representation of each word in the span). Then the model learns the weight of each feature based on the contexts, and generate new feature vectors $\varphi^*$:

$$a'_j = w_{a'} \cdot FFNN(\varphi_j f(x_u))$$
$$a'_{j,u} = \frac{\exp(a'_j)}{\sum_{j=1}^{V} \exp(a'_j)}$$
$$\varphi^* = \bigoplus_{j=1}^{V} a'_{j,u} \cdot \varphi_j$$

Where $\bigoplus$ is the concatenation operation and $f(x_u)$ is a linear function to map $x_u$ to the same dimension with the feature vector. $a'_{j,u}$ is the weight of each feature based on the contexts and $\varphi^*$ is the new reweighed feature vectors.

Span Feature Attention
To use features adaptively, we apply the Feature Attention mechanism to the span features: span width, grammatical number, and Metamap entity tags.

As shown in Figure 3, a new context-aware feature vector $\varphi^*$ is generated by the Feature Attention method and the new span features are applied to update the span representation, where $x^*_i$ is the contexts vectors generated by Bi-LSTM for span $i$ and $FA$ is the Feature Attention mechanism:

$$\varphi^*(i) = FA(\varphi(i), x^*_i)$$

$$g_i = [x^*_\text{START}(i), x^*_\text{END}(i), \hat{x}_i, \varphi^*(i)]$$

Coreference Score
The final coreference score of span $i$ and $j$ shows that (1) whether span $i$ is a mention, (2) whether span $j$ is a mention and (3) whether $j$ is an antecedent of $i$:

$$s(i,j) = \begin{cases} 
  s_m(i) + s_m(j) + s_a(i,j) + s_c(i,j), & j \neq \varepsilon \\
  0, & j = \varepsilon 
\end{cases}$$

where $s_m(i)$ is the mention score, $s_a(i,j)$ is the antecedent score, $s_c(i,j)$ is a rough sketch of likely antecedents and $w_c$ is a learned weight matrix.

Experiments
Dataset and Baseline
The experiments are performed on the BioNLP Protein coreference dataset [14]. For evaluation, we employed the scorer [2] provided by the organizers to make fair comparisons with previous work. We use the baselines below:

[7]: The system proposes a hybrid approach that combines both rule-based and learning-based method.
[4]: The system develops a rule-based system for anaphoric coreference resolution in the biomedical domain with simple modules derived from available systems.
[5]: The system designs a general modular framework, which is based on the smorgasbord architecture and contains multiple types of coreference types, and allows fine-grained specification of resolution strategies to resolve coreference.

[1]: https://metamap.nlm.nih.gov/MetaMapLite.shtml
[2]: http://bionlp-st.dbcls.jp/CO/eval-test/
Hyperparameters
We follow the same hyperparameters as in the [13]. We use a window size of 10 for the LSTM inputs. The maximum span width is 30 and only considers 50 antecedents. The model is trained up to 70 epochs with early stopping. We consider the following variations of our model:

BioNeu directly runs the [13] with integrating domain-specific information.

SFA (Span Feature Attention) uses Feature Attention mechanism to encode span feature $\varphi(i)$.

Results
Table 4 shows the performance comparison of our model with some baselines on the BioNLP development and test datasets. Overall, our SFA (span Feature Attention) model outperforms all baselines and achieves state-of-the-art performance with the 64.0% F1 score (+2.0 F1). In general, we notice that on the basis of ensuring the recall, our model performs much better precision than competitors, which indicates that after the distinct representation of mentions based on context, the noise brought by similar or identical mentions is reduced. Besides, we also find that the proposed model has a stronger generalization ability than competitors, which allows it to have better scalability.

Moreover, there are also some gaps in the performance of different models. First, BioNeu, similar to [9], makes improvement compared with other neural network-based baseline models. This indicates that domain-specific information pre-trained on large-scale corpora is practical in domain-related tasks when well-designed. Second, the proposed context-aware attention model can help the system to effectively represent mentions based on contexts and make more precise predictions, which can be supported by the comparison with other neural models ([6], [9], and BioNeu). To summarize, the results suggest that the distinction of mentions based on contexts is vital for effectively resolving coreference. In this case, context-aware attention models can help achieve this goal and make accurate predictions.

Discussion
Error Analysis
The scorer [3] provided by the organizer of the BioNLP shared task divides the results into three categories: 1) Gold: gold annotations of the dataset; 2) Predict: annotations predicted with two types: Correct and Spurious. Spurious means false annotations; 3) Missing: missing gold annotations. In order to better understand the performance of the proposed model, we analyzed these two types of errors in details: spurious link errors and missing link errors.

Figure 4 shows the detailed error statistics on the test dataset compared with some baselines ([6]-rule and [6]-neural). First, compared with [6], BioNeu performs much better in the reduction of spurious links, which indicates that the domain-specific information pre-trained on large-scale language models helps the model to predict more precisely. Second, compared with those neural network-based baselines [6]-neural and BioNeu, the proposed SFA model illustrates that the introduction of the Feature Attention mechanism further improves the model, greatly increasing the number of correct predictions and reducing the two errors. This shows that the distinction of the identical or similar mentions based on context can help the system learn more precisely and make more accurate predictions.

Identical Mention Linking Evaluation
Figure 5 and Figure 6 respectively display the performance of the BioNeu model and the SFA model on coreferential and incoferential identical mentions with different frequencies (the number of times that the identical mention appears in the document). As there are no coreferential identical mentions with a frequency greater than 3, we only show the frequency of 2 and 3. First, The performance of the SFA model on coreferential identical mentions is better than the BioNeu model. This indicates that the distinction of identical mentions helps in the prediction of the links between them to a certain extent. Second, for incoferential identical mentions, the SFA model performs better than the BioNeu model on all frequencies, and the superiority is significantly greater when the frequency of the identical mention is greater than 3. The superiority illustrates that, for the problem that the higher the frequency, the more difficult it is to predict, the Feature Attention mechanism does help to distinguish the identical mentions based on context, which provides further help in mention linking.

[1] http://bionlp-st.dbcls.jp/CO/eval-test/
Mention Detection Subtask
To further understand the utility of the Feature Attention mechanism for mention detection subtask, we list the mention detection performance in Table 5. Overall, compared with [9], the performance of the proposed SFA model is significantly improved by 6.3 F1 on the BioNLP development dataset. Moreover, the SFA model indeed performs much better than the BioNeu in the recall scores. This indicates that, in the baseline model, where there is a span not predicated as a mention, the other identical spans will likely not be detected as mentions due to their similar representations. However, in the SFA model, such false-negative errors are decreased, having benefited from the Feature Attention mechanism that reweights the features to distinguish identical spans with different representations based on different contexts.

Remaining Problems and Future Work
From the results of mention detection, we can find that the system is limited in the denoise of the context information when Feature Attention is added, which can be seen from the decrease in precision. In addition, the analysis of identical mention evaluation also show that the system gains less improvement when the frequency is less than 3. That means there are several potential improvements to our model as future work, such as how to adjust the setting of the FA so that the noise of the context information is reduced. Moreover, considering that this approach is still limited when context information is insufficient, we expect to utilize such information more fine-grained with the help of external knowledge base.

Case Study
To gain further insight into how identical terms’ representations can be distinguished by the attention mechanism, we take the former case in Figure 1 as an example to investigate the Feature Attention weights.

The attention weights of span features (span width, grammatical number, and Metamap entity tags) are shown in Figure 7. For the three ”it” in the figure, we can find that the first two ”it” that are coreferential gain similar weights for the three features, where the weight of span width is the highest, followed by the grammatical number, and finally Metamap. While, the third one has the opposite span weights to the first two with a higher weight in grammatical number than Metamap, though they are verbatim. This indicates that the identical spans that have the same features will have different span feature attention weights depending on their contexts through the span Feature Attention, thus improving the performance.

Conclusion
Identical mentions impose difficulties on the current methods of Bio-entity coreference resolution as they tend to get similar or even identical representations. This problem may directly lead to wrong predictions. In the paper, we focus on this issue and develop a context-aware attention model named Feature Attention and apply the attention mechanism in the process of span representation to adaptively exploit features and represent identical mentions considering different contexts. The results show that our model with the Feature Attention mechanism performs reasonably well in Bio-entity coreference resolution. The performance is supported on the BioNLP Protein Coreference dataset. Moreover, as our model learns to distinguish identical mentions more effectively, it achieves superior performance on the identical mention linking.

Acknowledgements
Not applicable.

Funding
This work has been supported by the National Key Research and Development Program of China (2018YFC0904004); National Natural Science Foundation of China (61772409); Innovative Research Group of the National Natural Science Foundation of China (61772102); Innovation Research Team of Ministry of Education (IRT_17R96); Project of China Knowledge Centre for Engineering Science and Technology.

Author details
1 School of Computer Science and Technology, Xi’an Jiaotong University, Xi’an, Shaanxi, P.R. China. 2 National Engineering Lab for Big Data Analytics, Xi’an Jiaotong University, Xi’an, Shaanxi, P.R. China. 3 Shaanxi Province Key Laboratory of Satellite and Terrestrial Network Technology Research and Development, Xi’an Jiaotong University, Xi’an, Shaanxi, P.R. China.

References
1. Dai, Z., Fei, H., Li, P.: Coreference aware representation learning for neural named entity recognition. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Macau, pp. 4946–4953 (2019).
Table 1 Statistics of identical mentions throughout the BioNLP dataset. Each represents its percentage of all mentions.

|                | Train (%) | Dev (%) |
|----------------|-----------|---------|
| Nonidentical   |           |         |
| Identical      |           |         |
| Coreferential  |           |         |
| Incoreferential|           |         |

Frequency

| Frequency | Train (%) | Dev (%) |
|-----------|-----------|---------|
| <10       | 22.0      | 24.3    |
| 10–100    | 13.7      | 21.6    |
| >100      | 64.3      | 54.1    |

Type

| Type       | Train (%) | Dev (%) |
|------------|-----------|---------|
| NP         | 18.3      | 13.9    |
| NN/NNS     | 5.6       | 2.1     |
| WDT        | 20.8      | 25.6    |
| PRP        | 7.6       | 7.2     |
| PRP$       | 16.6      | 18.4    |
| IN         | 27.1      | 28.6    |
| Others     | 4.0       | 4.2     |

Top-5

that, which, its, their, it

Table 2 Statistics of identical mentions in per document. Each represents its percentage of the total number of documents

|                | Train (%) | Dev (%) |
|----------------|-----------|---------|
| Without identical mentions | 57.1      | 57.3    |
| With identical mentions     | 42.9      | 42.7    |

Num

| Num | Train (%) | Dev (%) |
|-----|-----------|---------|
| 2   | 50.1      | 45.3    |
| 3   | 13.4      | 18.8    |
| 4   | 17.2      | 18.7    |
| 5   | 7.9       | 3.1     |
| >5  | 11.4      | 14.1    |

Table 3 Coreference resolution performance comparison by the average F1 value of three evaluation metrics.

| System         | Dataset | Dev-F1 | Test-F1 | Feature-based | Rule-based | Hybrid | Neural |
|----------------|---------|--------|--------|---------------|------------|--------|--------|
| BioNLP         | 62.4    | 60.9   | ✓      |               |            |        |        |
| BioNLP         | 66.6    |        |        |               |            |        |        |
| BioNLP         | 63.9    | 48.1   | ✓      |               |            |        |        |
| BioNLP         | 67.5    |        | ✓      |               |            |        |        |
| BioNLP         | 72.3    | 62.0   | ✓      |               |            |        |        |
| BioNLP, CRAFT  | 34.9    | 36.0   | 51.2   | 59.2          |            |        |        |
| CRAFT          | 50.5    | 65.7   | ✓      |               |            |        |        |
Table 4 The performance of protein coreference resolution with different models on two evaluation datasets of BioNLP.

|        | dev   |          | test  |          |
|--------|-------|----------|-------|----------|
|        | P     | R        | F1    | P        | R        | F1    |
| [7]    | 59.9  | 77.1     | 67.4  | 55.6     | 67.2     | 60.9  |
| [4]    | 63.4  | 64.4     | 63.9  | 46.3     | 50.0     | 48.1  |
| [5]    | 72.4  | 63.2     | 67.5  | /        | /        | /     |
| [6]-rule | 68.8  | 76.0     | 72.2  | 60.2     | 63.8     | 62.0  |
| [6]-neural | 60.4  | 61.9     | 61.2  | 54.9     | 58.0     | 56.4  |
| [9]    | 71.7  | 56.7     | 63.1  | 55.6     | 47.5     | 51.2  |
| BioNeu | 78.1  | 61.8     | 69.0  | 62.9     | 53.9     | 58.1  |
| SFA    | 71.9  | 65.8     | 68.7  | 69.0     | 59.5     | 64.0  |

Table 5 The performance of mention detection with different models on the development datasets of BioNLP.

|        | P     | R    | F1    |
|--------|-------|------|-------|
| [9]    | 82.0  | 66.3 | 73.3  |
| BioNeu | 84.1  | 73.1 | 78.2  |
| SFA    | 83.4  | 76.1 | 79.6  |

Figure 1 Example of linking errors of identical mentions affected by similar representations. Avoiding this error requires context-aware Feature Attention mechanism.

Figure 2 The Feature Attention model. The model learns to weigh each feature based on contexts.

Figure 3 The model of computing the span embedding representations.

Figure 4 Detailed error analysis compared with some baselines.

Figure 5 The performance of the two models on coreferential identical mentions with different frequencies within the document.

Figure 6 The performance of the two models on incoreferential identical mentions with different frequencies within the document.

Figure 7 An example of span Feature Attention weights in different features. Each row shows the attention weights of all the features of the span.