The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

Hierarchical Coherence Modeling for Document Quality Assessment

Dongliang Liao, Jin Xu[*], Gongfu Li, Yiru Wang

Data Quality Team, WeCaht, Tencent Inc., China.
{brightliao, jinxxu, gongfuli,dorisyrwang}@tencent.com

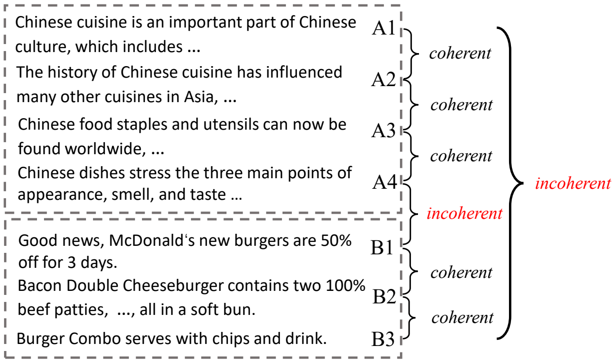
Abstract

Text coherence plays a key role in document quality assessment. Most existing text coherence methods only focus on the similarity of adjacent sentences. However, local coherence exists in sentences with broader contexts and diverse rhetoric relations, rather than just adjacent sentence similarity. Besides, the high-level text coherence is also an important aspect of document quality. To this end, we propose a hierarchical coherence model for document quality assessment. In our model, we implement the local attention mechanism to capture the location semantics, bilinear tensor layer to measure coherence and max-coherence pooling to acquire high-level coherence. We evaluate the proposed method on two realistic tasks: news quality judgement and automated essay scoring. Experimental results demonstrate the validity and superiority of our work.

Introduction

Document quality assessment is academically valuable and industrially applicable in many tasks, such as online news recommendation, automated essay scoring and readability assessment (Chen et al. 2010; Li and Hovy 2014; Dasgupta et al. 2018). Document quality is not only affected by the semantics, but also significantly affected by the text coherence (Petersen et al. 2015). Coherent documents are readable and attractive to readers, while documents with poor coherence are boring to read and may lead to misunderstanding. Several approaches are designed for modeling coherence, such as entity-based methods, lexical methods and neural coherence models. Entity-based methods connect by means of entities in sentences (Li and Hovy 2014; Nguyen and Joty 2017). Lexical methods measure the coherence of adjacent sentences based on word co-occurrence (Louis and Nenkova 2012; Chen et al. 2010). Neural coherence models learn vector representation of words and sentences, and capture local coherence by measuring the similarity of words or sentences vectors (Mesgar and Strube 2018; Moon et al. 2019).

However, most existing local coherence models only focus on the similarity of adjacent sentences, which is inadequate for modeling the full coherence of documents. Two attributes of text coherence are ignored in existing methods.



Chinese cuisine is an important part of Chinese culture, which includes ...

The history of Chinese cuisine has influenced many other cuisines in Asia, ...

Chinese food staples and utensils can now be found worldwide, ...

Chinese dishes stress the three main points of appearance, smell, and taste ...

Good news, McDonald's new burgers are 50% off for 3 days.

Bacon Double Cheeseburger contains two 100% beef patties, ..., all in a soft bun.

Burger Combo serves with chips and drink.

Figure 1: An example of high-level coherence.

Firstly, there are various rhetorical relations between sentence and the broader context, rather than the similarity of adjacent sentences. For example, in parallel structures, several consecutive sentences form a coherent and consistent discourse unit, but are relative independent with each other. An example of this circumstance is shown as follows:

A: *Let's see the new features in IPhone 11.*
B: *It's powered by the latest proprietary chip, the A13 Bionic processor.*
C: *New tri-lens camera array combines a 12-megapixel main lens, an ultra-wide angle lens and a telephoto lens.*
D: *The iPhone 11 also features a new anodized aluminum finish, which Apple says is more durable.*

where sentences A,B,C and D form a coherent part about the new features of IPhone 11, but each sentence describes an independent feature. Therefore, the first challenge is **how to capture the various rhetorical relation between sentences and broader contexts**.

Secondly, the document quality is not only impacted by sentence-level coherence, but also hierarchical coherence of high-level discourse units. The high-level coherence reflects the organizational structure and logicality. Figure 1 shows an example of this circumstance, where the first four sentences are about Chinese cuisine, while the next three sentences form an advertisement. The inserted advertisements hurt the reading experience and decrease document quality. As for the sentence-level coherence, only A4 and B1 are incoherent, but the overall text is incoherent. Thus the second challenge is **how to model the hierarchical coherence of**

[*] Corresponding author.

documents. Discourse analysis is a related topic focused on studying text coherence and structure. Unfortunately, the reliance on labeled data and the error propagation are obstacles to the usage and performance of discourse parsers in document modeling (Liu and Lapata 2018).

Our motivation is to overcome these limitations, inducing document representations and coherence representations directly by end-to-end neural networks. First of all, we build the sentence representation and consider the sentences as the basic discourse units. Then we decompose the hierarchical coherence modeling into three steps: merging low-level discourse units to local context block, capturing the coherence of each local context block and detecting the coherent block as high-level discourse units. This process can be stacked hierarchically to acquire multi-level discourse units and hierarchical coherence. Concretely, we utilize Transformer encoder to build the sentence vectors. We combine the convolutional operation and attention mechanism to generate local context block of low-level discourse units. We employ bilinear layers to capture multiple dimension rhetoric relations of each unit of local context, and merge all units coherence as the context coherence. Then, we propose max-coherence pooling to to detect the coherent context blocks as high-level discourse units. At last, we aggregate all the coherence vectors of all discourse units in different levels by attention pooling, representing the overall coherence of document. We summarize our contributions as follows.

- We capture the various rhetorical relation in a broader context rather than the similarity of adjacent sentences for modeling the text coherence, and introduce the hierarchical coherence for document quality assessment.
- We propose the local attention pooling to generate local context block, capture multiple dimension coherence by bilinear layer and propose max-coherence pooling to to detect high-level discourse units.
- We evaluate the proposed method on two realistic tasks: online news quality judgement and automated essay scoring. The experimental results demonstrating the validity and superiority of our approaches [1].

Related Work

Coherence Modeling

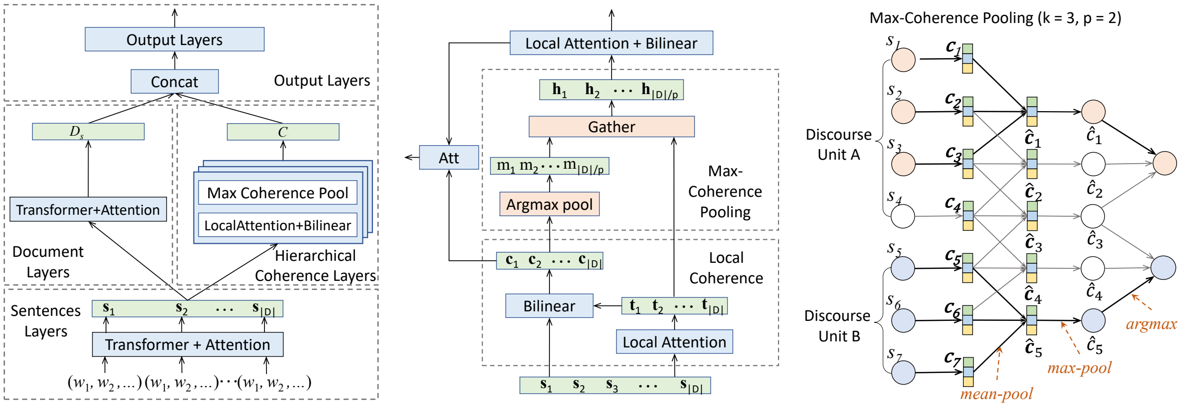
Text coherence modeling has attracted many researchers' interest. Early works leveraged lexical and syntax features for capturing coherence. Barzilay and Lapata (2008) proposed the entity-based model, extracting entities from sentences and measuring coherence by the occurrence frequency of entities. Nguyen and Joty (2017) deployed CNN on entity grids to improve the coherence modeling performance. However, entity based models rely on lexical and syntax analysis for entity extraction, which may encounter error propagation. Recently, end-to-end neural coherence models got state-of-the-art performance in related tasks. Li and Hovy (2014) proposed a neural method, representing sentences with recurrent and recursive neural network, and estimating the coherence probability with local coupled layers for the window of three sentences. Nadeem and Ostendorf (2018) proposed a hierarchical RNN with a bidirectional context for capturing the sentence coherence with their adjacent sentences. Han et al. Moon et al. (2019) captured text coherence by bi-linear layer for discourse relations and light weight convolution-pooling for the attention and topic structures. All these methods only modeled the similarity of adjacent sentences. Our approaches capture multi-dimensional coherence in broader contexts and supplement the high-level coherence for better understanding the text coherence.

Our work is partly motivated by another related research topic, *discourse analysis*. Rhetorical Structure Theory is the most widely accepted discourse structure, providing a systematic way to analyze the text (Marcu 2000; Taboada and Mann 2006). Scholars have developed automated discourse parsers (Ji and Eisenstein 2014; Chuan-An et al. 2018) for generating RST style trees from documents, facilitating some applications such as the text classification (Ji and Smith 2017) and sentiment analysis (Bhatia, Ji, and Eisenstein 2015). However, discourse parsing is such a difficult task and there is still a large margin to construct a perfect parser (Chuan-An et al. 2018). Errors in discourse tree significantly impact the document modeling performance. In this paper, we do not try to detect the discourse structure precisely, but only focus on capturing the hierarchical coherence of documents. We propose an approximate but valid way to aggregate coherent consecutive sentences a high-level discourse units, which is efficient and can be implemented in end-to-end neural networks.

Document Quality Assessment

Two typical categories of document quality assessment tasks are automated essay scoring (AES) (Dasgupta et al. 2018; Alikaniotis, Yannakoudakis, and Rei 2016) and readability assessment (Li and Hovy 2014; Petersen et al. 2015; Todirascu et al. 2016). Early studies mainly focused on feature engineering such as bag of words, N-gram and other linguistics features (Chen et al. 2010; Feng et al. 2010; Todirascu et al. 2016). Up to now, many scholars followed the popular neural approaches for document quality assessment, such as CNN, RNN and attention mechanism (Dong, Zhang, and Yang 2017; Dasgupta et al. 2018; Moon et al. 2019). However, these methods failed to capture the discourse coherence, which significantly influenced the document quality. Louis and Nenkova (2012) and (Miltsakaki and Kukich 2004) showed that discourse coherence features significantly strengthened the readability assessment and essay scoring performance. Mesgar and Strube (2018) proposed an end-to-end neural coherence model which captures coherence vectors of adjacent sentences and captures the text coherence by convolutional operators on the coherence vector. Nadeem and Ostendorf (2018) and Tay et al. (2018) also proposed their neural coherence methods for readability assessment and essay scoring, with CNN and SkipFlow RNN for coherence modeling. Comparing with these existing approaches, we proposed a novel coherence model for capturing text coherence, achieving considerable performance improvement in document quality assessment tasks.

[1] Code and Dataset: https://github.com/BrightLiao/HierCoh


Output Layers
Concat
Output Layers
D_s
C
Transformer+Attention
Max Coherence Pool
LocalAttention+Bilinear
Document Layers
Hierarchical Coherence Layers
Sentences Layers
$\mathbf{s}_1$ $\mathbf{s}_2$ ⋯ $\mathbf{s}_{|D|}$
Transformer + Attention
$(w_1, w_2, \ldots)(w_1, w_2, \ldots)\cdots(w_1, w_2, \ldots)$
Local Attention + Bilinear
$\mathbf{h}_1$ $\mathbf{h}_2$ ⋯ $\mathbf{h}_{|D|/p}$
Gather
Att
m_1 $\mathrm{m}_2 \cdots \mathrm{m}_{|D|/p}$
Argmax pool
Max-Coherence Pooling
$\mathbf{c}_1$ $\mathbf{c}_2$ ⋯ $\mathbf{c}_{|D|}$
Local Coherence
Bilinear
$\mathbf{t}_1$ $\mathbf{t}_2$ ⋯ $\mathbf{t}_{|D|}$
Local Attention
$\mathbf{s}_1$ $\mathbf{s}_2$ $\mathbf{s}_3$ ⋯ $\mathbf{s}_{|D|}$
Max-Coherence Pooling (k = 3, p = 2)
Discourse Unit A
Discourse Unit B
s_1 s_2 s_3 s_4 s_5 s_6 s_7
c_1 c_2 c_3 c_4 c_5 c_6 c_7
$\hat{\mathbf{c}}_1$ $\hat{\mathbf{c}}_2$ $\hat{\mathbf{c}}_3$ $\hat{\mathbf{c}}_4$ $\hat{\mathbf{c}}_5$
$\hat{c}_1$ $\hat{c}_2$ $\hat{c}_3$ $\hat{c}_4$ $\hat{c}_5$
argmax
max-pool
mean-pool

(a) Framework of HierCoh model.

(b) Coherence Layers.

(c) An example of MCP.

Figure 2: (a) Framework of the proposed HierCoh model. (b) Illustration of coherence layers, consisting of local coherence and max-coherence pooling. (c) Illustration of max-coherence pooling process on the example of Fig.1. Suppose $\mathbf{c}_i$ represents the local coherence vector of sentence s_i generated by Eq. (6), we calculate $\hat{c}_i$ with Eq. (8) and select the biggest $\hat{c}_i$ to form the approximate high-level discourse unit.

Model

We propose a novel coherence model named HierCoh, shown in Fig. 2(a). We implement a hierarchical architecture, consisting of sentence layers, hierarchical coherence layers, document layers and output layers.

Let $\mathcal{D}$ represent a document, consisting of a sentences sequence $\mathcal{D} = (s_1, s_2, \cdots, s_{|\mathcal{D}|})$, where each sentence is a sequence of one-hot words $s = (w_1, w_2, \cdots, w_{|s|})$. Sentence layers take each sentence s as the input and generate the sentence vector **s**. Hierarchical coherence layers capture the hierarchical text coherence as a vector C. Document layers aggregate sentence vectors as document semantics vector D_s and integrate D_s coherence vector C for generating document vector D. Then document vector D is fed into task-specific output layers for document quality assessment.

Sentence Layers

As mentioned before, we treat the sentences as the basic discourse units. In sentence layers, we embed sentences to represent vectors. We map one-hot word vectors to dense embeddings and utilize Transformer encoder (Vaswani et al. 2017) to update word vectors with the word context, then deploy attention pooling for aggregating word vectors to sentence vectors.

Hierarchical Coherence Layers

In Hierarchical coherence layers, we generate the local context block of low-level discourse units, capture the multi-dimensional coherence, and detect high-level discourse units for capturing the hierarchical coherence.

Local Context Block Convolution neural network (CNN) is a powerful tool for modeling local structures in computer vision tasks (Zeiler and Fergus 2014). The fixed-weight filters in CNN are designed for capturing the contour or texture of image pixels. The stacked CNN can extract hierarchical structures efficiently.

However, the relations between sentence semantics are more complicated and flexible. In this work, we integrate the convolution operation with an attention mechanism, named Local Attention Pooling (LAP), for tackling this challenge. LAP replaces the linear transformation in CNN to the multi-head attention pooling. Attention pooling generates attentive weights α_i depending to the input sequence S_i, and merges the input sequence S_i by multiplying attentive weights α_i. Multi-head attention is a extension of attention mechanism, which calculates multiple attentive weights to construct multiple attention "heads". The output of multi-head attention is concatenation of all attention "heads". The calculation of LAP can be formulated as follows, where W_a^m, V_a^m are parametric matrix.

$$\mathbf{v}_j^m = \tanh(W_a^m \mathbf{s}_j) \tag{1}$$

$$\alpha_j^m = \frac{\exp\left(V_a^m \mathbf{v}_j^m\right)}{\sum_{i-k/2}^{i+k/2} \exp\left(V_a^m \mathbf{v}_t^m\right)} \tag{2}$$

$$\text{head}_i^m = \sum_{i-k/2}^{i+k/2} \alpha_j^m \mathbf{s}_j \tag{3}$$

$$\mathbf{t}_i = Lap^k\left(\mathbf{s}_{i-\frac{k}{2}}, \cdots, \mathbf{s}_i, \cdots, \mathbf{s}_{i+\frac{k}{2}}\right) \tag{4}$$

$$= \text{Concat}\left(\text{head}_i^1, ..., \text{head}_i^m\right) \tag{5}$$

Comparing with the fixed-weight linear transformation in CNN, the multi-head attention pooling in LAP aggregates the context semantics with a flexible weights depending on the semantics of each sentence. We treat $\mathbf{t}_i$ as the local context block of sentences $\mathbf{s}_{i-\frac{k}{2}}, \cdots, \mathbf{s}_i, \cdots, \mathbf{s}_{i+\frac{k}{2}}$.

Multi-Dimension Coherence Given the context vector $\mathbf{t}_i$ of sentence $\mathbf{s}_i$, an intuitive way to estimate the local coherence is using vector similarity such as the cosine similarity. As we have explained, the relation between sentences and their context block is not only similarity but diverse rhetoric relations (Taboada and Mann 2006). Inspired the neural tensor layer (Socher et al. 2013; Tay et al. 2018), we implement a bilinear tensor layer to model the relations of the two vectors across multiple dimensions.

$$\mathbf{c}_i = \sigma(\mathbf{s}_i^T \mathbf{W}_b \mathbf{t}_i + \mathbf{b}) \tag{6}$$

Here, $\mathbf{W}_b \in \mathbb{R}^{d_s \times d_c \times d_t}$ is called the relation tensor, and $\mathbf{b} \in \mathbb{R}^{d_c}$ is the bias vector. d_s, d_t are the dimension of sentence vector **s** and context vector **t**. σ denotes sigmoid function. d_c is the dimension of output coherence vector $\mathbf{c}_i$.

We consider each slice $W_b^{(r)} \in \mathbb{R}^{d_s \times d_t}$ in $\mathbf{W}_b$ represents a kind of implicit and learnable rhetorical relation r. Thus, each element $c_i^{(r)} = \sigma(\mathbf{s}_i^T W_b^{(r)} \mathbf{t}_i)$ in $\mathbf{c}_i$ represents the probability of the relation r between sentence $\mathbf{s}_i$ and local context block $\mathbf{t}_i$. With the bilinear tensor layer, we can extract the local coherence of all sentences as $C^{(0)} = \left[\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_{|\mathcal{D}|}\right]$.

Hierarchical Coherence In this work, we propose an approximate approach, called Max-Coherence Pooling (MCP), leveraging low-level coherence (i.e. level $l-1$) to detect high-level discourse units for measuring the high-level coherence (i.e. level l). The behind intuition is that if several consecutive low-level discourse units are in coherence, they can be regarded as a coherent high-level discourse unit.

Concretely, after we extract the context block and local coherence of low-level discourse units with LAP and bilinear tensor layers, MCP can be divided into three steps. Firstly, we exploit the mean pooling on coherence vectors $\mathbf{c}_{i-\frac{k}{2}}^{(l-1)}, \cdots, \mathbf{c}_i^{(l-1)}, \cdots, \mathbf{c}_{i+\frac{k}{2}}^{(l-1)}$ in the local context block for the overall multi-dimension coherence $\hat{\mathbf{c}}_i^{(l-1)}$. Secondly we treat the max dimension $\hat{c}_i^{(l-1)}$ in $\hat{\mathbf{c}}_i^{(l-1)}$ as the coherence probability of the local context block $\mathbf{t}_i^{(l-1)}$. Thirdly we adopt a "argmax pooling" to select the low-level context m_i with max coherence score $\hat{c}_i$ in a k size window with strides p.

$$\hat{\mathbf{c}}_i = \frac{1}{k}\left(\mathbf{c}_{i-\frac{k}{2}} + \cdots + \mathbf{c}_i + \cdots + \mathbf{c}_{i+\frac{k}{2}}\right) \tag{7}$$

$$\hat{c}_i = \max\left\{c_{ij} \middle| c_{ij} \in \hat{\mathbf{c}}_i\right\} \tag{8}$$

$$m_i = \underset{j}{\operatorname{argmax}}\left\{\hat{c}_j \middle| j \in [i - \frac{k}{2}, i + \frac{k}{2})\right\} \tag{9}$$

By these two steps, we have selected the consecutive coherent low-level discourse units in window size k, which is the local context block $\mathbf{t}_{m_i}^{(l)}$. Thus we treat them as the approximation of a high-level discourse unit.

$$\mathbf{h}_i^{(l)} = \text{MCP}(\mathbf{h}_i^{(l-1)}) = \mathbf{t}_{m_i}^{(l-1)} \tag{10}$$

Context window size k and strides p are hyper-parameter, controlling the coherence scope and selection rate of max-coherence pooling. Figure 2(b) shows the illustration of the coherence layer. Fig. 2(c) presents max-coherence pooling process on the example of Fig. 1.

Then, given the high-level discourse units, we can also adopt the local attention pooling and bilinear tensor layer to measure the high-level coherence. Our model can capture the hierarchical coherence by stacking multi-layer max-coherence pooling, local attention pooling and bilinear layers. At last, we adopt attention pooling to aggregate all coherence vectors in different levels to the document coherence vector C. The entire process is shown in Alg. 1, where L is the coherence layer number.

Algorithm 1: Text Coherence Modeling

```
Input: $L, \{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_{|\mathcal{D}|}\}$
Result: Text coherence vector $C$
$H^{(0)} \leftarrow \{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_{|\mathcal{D}|}\}$;
Generate $T^{(0)} = \{\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_{|\mathcal{D}|}\}$ by (5);
Generate $C^{(0)} = \{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_{|\mathcal{D}|}\}$ by (6);
$\hat{C} \leftarrow C^{(0)}$;
for $l \in [1, L]$ do
    Generate $H^{(l)}$ by (7)(8)(9)(10);
    Generate $T^{(l)}$ by (5);
    Generate $C^{(l)}$ by (6);
    $\hat{C} \leftarrow \hat{C} \cup C^{(l)}$;
end
$C \leftarrow AttentionPool(\hat{C})$
```

Document Layers

Document quality is determined by both the text coherence and the semantics. We integrate the all the sentence vectors by another Transformer and attention pooling layer as the document semantics vector D_s for document semantics modeling. We concatenate document semantics vector D_s and coherence vectors C as the final document vector $D = [D_s, C]$, so that D captures both the semantics and hierarchical coherence of document. Then D can be fed into output layers for document quality assessment.

Evaluate Tasks

Automated Essay Scoring

Automated essay scoring (AES) is a classic document quality assessment problem, since the score of student essay is determined by the essay quality. The most popular AES dataset is Automated Student Assessment Prize (APSP) corpus. In total, ASAP consists of eight sets of essays, each of which associates with one prompt. These essays were originally written by students between Grade 7 and Grade 10. The statistics information of APSP please refers to Taghipour and Ng (2016). AES is usually considered as a regression task to estimate the essay score. Each essay is scored by domain experts, and score ranges of different essay sets are scaled to $[0, 1]$. We implement a sigmoid output layer, taking document vector D as the input to get the predicted essay scores $\hat{y}$, and mean square errors are employed as the loss function.

Online News Quality Judgement

We collect a Chinese online news dataset from WeChat [2] for document quality assessment. Similar as some readability assessment datasets such as De Clercq et al. (2014), we

[2] https://www.wechat.com/

Category	#Pairs	1st. News Len.	2nd. New Len.
0	4214	515	522
1	3908	642	427
2	3840	476	589

Table 1: Statistics of Online News Dataset.

define the quality assessment task as a pair-wise rank task, judging which one has better quality in a news pair. We collect two news describing the same topic or event as a news pair and employ human annotators to annotate the pair-wise quality rank labels of three categories: $\{0, 1, 2\}$. Category 0 indicates that the news pair has equal quality, category 1 represents the first news has better quality and category 2 denotes the second news has better quality. The quality labels follow the criteria of *clear theme*, *coherent sentences* and *well organized structure*. We employed 8 annotators and 1 checker for the Chinese news dataset. We shuffled all the news pairs into batches and handed out to 8 annotators. The checker is an experienced annotator who also participated in the development of annotating criteria. The checker randomly checked 20% of each batch by annotating the news pair himself and comparing it with the results of annotators. A batch with more than 15% divergence in check examples would be rejected and re-annotated. At last, we got 11962 valid news pairs. The dataset statistics information is shown in Table 1.

The online news quality judgement is a three-classification problem. We adopt a Siamese architecture for modeling the news pair and get two document vectors D_1 and D_2. We concatenate the document vectors together and adopt a softmax output layer to predict the probability of news pair's labels. The cross entropy loss function is employed as the minimized object.

Experiments

Baselines

For the AES task, we report following baselines.

- **CNN+LSTM**. Taghipour and Ng (2016) is A neural method for the AES task, ensembling CNN and LSTM for essay scoring.
- **LSTM-CNN-att**. Dong, Zhang, and Yang (2017) proposed a hierarchical structure with attention CNN for sentence modeling and LSTM for document modeling.
- **SkipFlow**. Tay et al. (2018) adopted tensor compositions to model the text coherence for essay scoring.

For the online news quality judgement task, we employ the state-of-the-art coherence models.

- **CNNCoh**. Farag, Yannakoudakis, and Briscoe (2018) adopt CNN to capture local coherence of sentences as a coherence score.
- **CohLSTM**. Mesgar and Strube (2018) proposed a model to encoder the perceived coherence of a text by a vector, named CohLSTM. We integrate CohLSTM with HAN to compare CohLSTM with HierCoh.
- **UNCM**. Moon et al. (2019) proposed a unified neural coherence model for local and global coherence discrimination, implementing BiLSTM, bilinear and CNN layers.

Besides the task-specific methods, we take several commonly used document modeling methods for comparing.

- **HAN**. The hierarchical attention network (Yang et al. 2016) adopts 2-layer BiGRU and attention mechanism for document modeling.
- **Bert**. We finetune the pretrained Bert-base, taking the [CLS] vector of the last layer as document representations for document quality assessment (Devlin et al. 2019).
- **ToBert**. Transformer over Bert is a hierarchical architecture, encoding sentences with Bert and merging sentence vectors with Transformer (Pappagari et al. 2019; Wang et al. 2020).

We also report the simplified version of the proposed method for ablation analysis.

- **H-Trans**. The hierarchical network adopts staked Transformer and attention pooling layers to generated document vectors without coherence modeling.
- **H-Trans+Cos**. We replace bilinear tensor layers with cosine similarity for coherence modeling.
- **H-Trans+LC**. We integrate local coherence with H-Trans, without MCP layers.
- **OnlyMCP**. We model the document only using the coherence vector C, without document layers.

Implementation Settings

For the AES task, The hidden sizes of H-Trans and proposed methods are empirically set as 64 for word embedding, Transformers and attention layers. For the news quality judgement task, the hidden sizes are set as 128. In coherence layers, there are several hyper-parameters that need to be set. We set the coherence vector size (i.e. the hidden size of bilinear layer) as 5 follows Tay et al. (2018). The window size k and layer number of max-coherence pooling L is fine tuned on $\{3, 5, 7, 11\}$ and $\{1, 2, 4, 8\}$ respectively. The strides p is set as the half of window size $p = k/2$ empirically. We adopt the Adam with 0.0005 learning rate for training and employ a dropout mechanism on the input word embedding with dropout rate 0.5. All experiments are constructed based on TensorFlow with Tesla P40 GPU.

Automatic Essay Scoring Result

In the following experiments, we follow the 5-fold evaluation method with Taghipour and Ng (2016) [3] and reuse the data preprocess code of Dong, Zhang, and Yang (2017) [4]. The metrics are quadratic weighted Cohen's κ (QWK).

Table 2. shows the overall performance comparison of AES. We can observe that the proposed HierCoh achieves the best results in essay sets 1, 2 and 7, and enhances the QWK in set 3 and 6 to approach the Bert-based models performance. However, our methods perform poorly on essay

[3] https://github.com/nusnlp/nea/tree/master/data

[4] https://github.com/feidong1991/aes/

Models	#Params	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Avg. 1-8	Avg 1-7
CNN+LSTM §	–	0.821	0.688	0.694	0.805	0.807	0.819	0.808	0.644	0.761	0.777
LSTM-CNN-att §	326K	0.822	0.682	0.672	0.814	0.803	0.811	0.801	0.705	0.764	0.772
SkipFlow §	–	0.832	0.684	0.695	0.788	**0.815**	0.810	0.800	0.697	0.764	0.775
HAN†	449K	0.825	0.693	0.691	0.798	0.798	0.808	0.809	0.637	0.757	0.775
Bert†	110M	0.821	0.678	**0.717**	**0.815**	0.803	0.813	0.802	**0.718**	**0.772**	0.780
ToBert†	119M	0.823	0.702	0.711	0.806	0.808	**0.832**	0.806	0.657	0.769	0.784
H-Trans	363K	0.818	0.699	0.703	0.788	0.798	0.806	0.811	0.639	0.758	0.775
H-Trans+Cos$_{k=5,L=1}$	367K	0.827	0.699	0.707	0.792	0.798	0.821	0.812	0.637	0.762	0.778
H-Trans+Cos$_{k=5,L=4}$	370K	0.829	0.700	0.713	0.807	0.804	0.816	0.792	0.647	0.764	0.781
H-Trans+LC$_{k=5,L=1}$	385K	0.837	0.696	0.710	0.798	0.812	0.826	0.799	0.633	0.763	0.782
OnlyMCP$_{k=5,L=4}$	455K	0.828	0.684	0.699	0.775	0.803	0.819	0.771	0.594	0.747	0.768
HierCoh$_{k=5,L=4}$	464K	**0.839**	**0.702**	0.711	0.809	0.801	0.827	**0.820**	0.631	0.763	**0.786**

Table 2: Automated essay scoring performance. We sign the best result in bold and second best result with underlines. Methods with up-script § is the result reported in related works. Methods with up-script † is our reproduced version.

set 8, so that the average QWK on all essay sets is similar with LSTM-CNN-att and SkipFlow, obviously worse than Bert-based methods. That's because the set 8 has the longest essays (with average length of 650) but least examples (only 723 in total), while other sets have more than 1500 essays and average lengths are 150-350.

Modeling the document only with the coherence vector (i.e. OnlyMCP) gets the worst performance on the most essay sets. It is not surprising since the scores of student essays are not only depend on the text coherence, but also the semantic features such as their topics and arguments. HAN, LSTM-CNN-att and H-Trans have similar hierarchical structures but different neural units in sentences and document layers. HAN and H-Trans perform slightly better on set 1-7, while LSTM-CNN-att is much better on set 8 with simpler layers and fine-tuning parameters. H-Trans has less parameters than HAN and much faster training speed with parallelizable Transformer architecture, so we adopt H-Trans to model the document semantics.

H-Trans+Cos$_{k=5,L=1}$ and H-Trans+Cos$_{k=5,L=4}$ enhance the QWK obviously on set 1-7, with slightly parameters increment of LAP. LAP has the same parameter sharing characteristics as CNN, extracting useful features with a small amount of parameters. To some extent, we can treat the deep neural layers as the feature extractors. The coherence layers are heuristic high-order feature interactions of sentence semantic features. Cosine similarity can be seen as a kind of non-parametric feature interaction, increasing no model complexity of feature extract layers but supplementing effective features. In this perspective, bilinear tensor layer is a parametric feature interaction, generating more dimensional, more effective and more flexible feature interactions. Thus H-Trans+LC$_{k=5,L=1}$ and HierCoh$_{k=5,L=4}$ further improve the performance on set 1-7. Nevertheless, bilinear tensor layers also increase the model complexity, making it harder to train the model well on set 8. Comparing with Bert-based methods, the proposed methods have much less parameters and achieve similar performances, demonstrating that not only the complex model but also the reasonable architecture innovations brings the improvement.

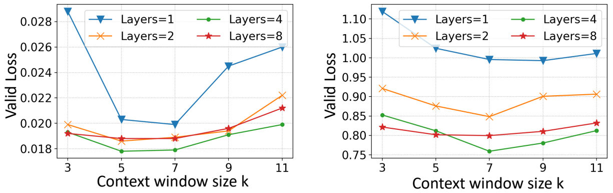
Models	Acc	0-F1	1-F1	2-F1
CNNCoh†	0.576	0.516	0.586	0.602
CohLSTM†	0.572	0.404	0.625	0.614
UNCM†	0.547	0.465	0.506	0.551
HAN†	0.551	0.505	0.549	0.591
Bert†	0.596	0.545	0.629	0.643
ToBert†	0.607	**0.565**	0.627	0.651
H-Trans	0.553	0.517	0.559	0.586
H-Trans+Cos$_{k=7,L=1}$	0.595	0.481	0.595	0.619
H-Trans+Cos$_{k=7,L=4}$	0.625	0.507	0.621	0.661
H-Trans+LC$_{k=7,L=1}$	0.621	0.489	0.644	0.649
OnlyMCP$_{k=7,L=4}$	0.611	0.487	0.617	0.634
HierCoh$_{k=7,L=4}$	**0.644**	0.532	**0.687**	**0.688**

Table 3: News quality judgement comparison. We sign the best result in bold. Methods with up-script † is our reproduced version.

Online News Quality Judgement Result

We sample 80% of news pairs as the training set, 10% news pairs as the validation set and 10% as the test set. We measure the online news quality judgement task by classification accuracy and F1 score of each category. For a category c, we treat the other two categories of examples as negative examples to calculate the F1 score. The overall prediction performance is shown in Table 3. The proposed approach HierCoh$_{k=7,L=4}$ achieves the best performance.

Among all baselines, UNCM gets the worst result. This may be because the UNCM is designed for the discrimination of sentence order, instead of the overall document quality. Different from AES tasks, existing document modeling methods (i.e. HAN, Bert and ToBert) perform poorly in news quality judgement. These methods are effective to capture the semantics of documents, while the news pair is associated with the same topic or events, having similar semantics. Coherence features play a more important role in the quality judgement in this situation. CNNCoh and CohLSTM capture the coherence between sentences and their adjacent sentences, thus integrating them with H-Tran improves the news quality judgement capability. In this work, we argue that the coherence between in the broader



(a) AES (b) News Quality Judgement.

Figure 3: Valid losses with different context window size and coherence layers.

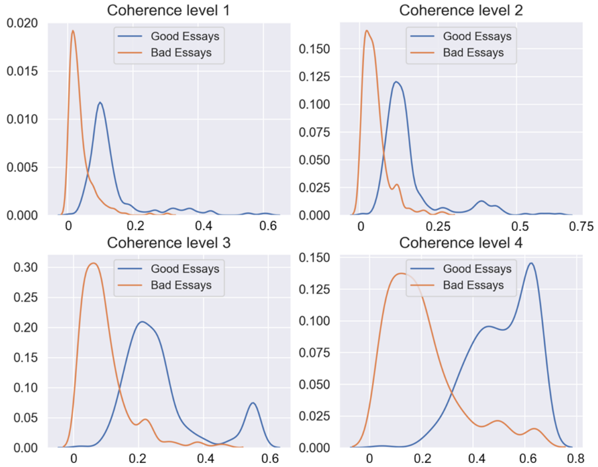
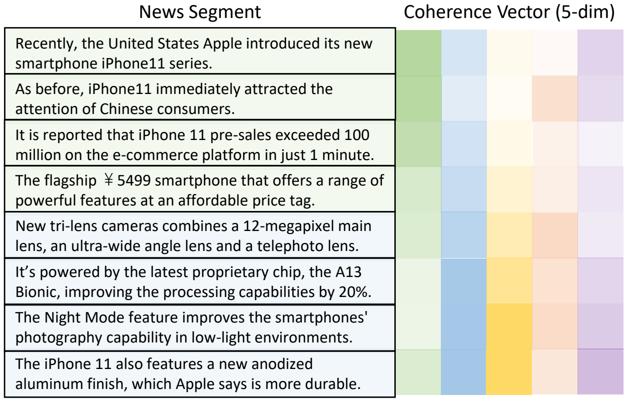


Figure 4: Coherence score distribution.

context should be considered, rather than just the similarity of adjacent sentences. Thus the H-Tran+Cos$_{k=7,L=1}$ and H-Tran+LC$_{k=7,L=1}$ perform better than existing coherence models. What's more, we capture hierarchical coherence of document by max-coherence pooling in HierCoh, further improve the document quality judgement performance significantly. Note that OnlyMCP method also performs a considerable result, demonstrating the superiority of capturing the hierarchical multi-dimensional coherence.

Further Analysis

Parameters Effects There are two important hyperparameters in the HierCoh models, context window size k in Eq. (5) and max coherence layers in Alg. 1. Figure 3 shows the average valid loss variant on two tasks with $k \in \{3, 5, 7, 11\}$ and $L \in \{1, 2, 4, 8\}$. Model with coherence layer number $L = 1$ local size $k = 3$ is similar to previous coherence models that only capture the local coherence with adjacent sentences, which is insufficient for document quality assessment. Models with multiple coherence layers of coherence modeling achieve lower valid loss. Actually, the coherence between sentences and their broader context can also be captured in high-level coherence modeling, so that models with multiple coherence layers still achieve an appreciable performance improvement with local size $k = 3$. Even though, local size k still affects the impact scope of high-level coherence. Comparing the validation loss, we assign local size $k = 5$ and coherence layer number $L = 4$ for automated essay scoring. The valid loss curve of news quality judgement shows slightly different with AES, achieving best loss when $k = 7$, since the document length is longer in the online news dataset. Thus we set $k = 7$ and $L = 4$ for the news dataset.



Recently, the United States Apple introduced its new smartphone iPhone11 series.

As before, iPhone11 immediately attracted the attention of Chinese consumers.

It is reported that iPhone 11 pre-sales exceeded 100 million on the e-commerce platform in just 1 minute.

The flagship ￥5499 smartphone that offers a range of powerful features at an affordable price tag.

New tri-lens cameras combines a 12-megapixel main lens, an ultra-wide angle lens and a telephoto lens.

It's powered by the latest proprietary chip, the A13 Bionic, improving the processing capabilities by 20%.

The Night Mode feature improves the smartphones' photography capability in low-light environments.

The iPhone 11 also features a new anodized aluminum finish, which Apple says is more durable.

Figure 5: Coherence Vector Visualization.

Hierarchical Coherence Distribution For illustrating the effect of hierarchical coherence, we output coherence vectors $C^{(l)}$ of in different level of student essays. We calculate the coherence probability $\hat{c}$ by Eq.(8) in level l of document $\mathcal{D}$. We select the high score essays as *good essays* (normalized score > 0.8) and low score essays (normalized score < 0.3) as *bad essays* to observe the coherence score distribution of each level by kernel density estimation, illustrated in Fig 4. This observation demonstrates that the high-level coherence takes effect in document quality assessment. What's more, the high-level coherence scores have more stable distribution and less outliers, which makes the high-level coherence distribution easier to be captured by neural models.

Case Study of Multi-dimension Coherence Figure 5 shows the illustrating of the coherence vectors of a translated news segment in online news datasets. In the proposed model, each dimension of coherence vector reflects the probability of a kind of relationship between discourse units. We use different colors to denote different dimensions of coherence vectors. The deeper the color denotes the bigger the value. This news segment is the first paragraph of an online news about Apple smartphones, from where we detect the parallel structure example in Introduce section. The first four sentences are consequent relations and the latter four sentences are coordinate/parallel structure. We can observe that the green dimension reflects the consequent relation while the blue and yellow dimensions capture the parallel structure. This observation proves that bilinear layers can capture various relations in multi-dimensional coherence vectors.

Conclusion

In this paper, we argue that text coherence exists in a broader local context and between high-level discourse units. Based on that, we develop a hierarchical coherence model for document quality assessment. Experiment results on two realistic tasks demonstrate the superiority of our method.

References

Alikaniotis, D.; Yannakoudakis, H.; and Rei, M. 2016. Automatic text scoring using neural networks. *arXiv preprint arXiv:1606.04289* .

Barzilay, R.; and Lapata, M. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics* 34(1): 1–34.

Bhatia, P.; Ji, Y.; and Eisenstein, J. 2015. Better document-level sentiment analysis from rst discourse parsing. *arXiv preprint arXiv:1509.01599* .

Chen, Y.-Y.; Liu, C.-L.; Lee, C.-H.; Chang, T.-H.; et al. 2010. An unsupervised automated essay-scoring system. *IEEE Intelligent systems* 25(5): 61–67.

Chuan-An, L.; Huang, H.-H.; Chen, Z.-Y.; and Chen, H.-H. 2018. A unified RvNN framework for end-to-end chinese discourse parsing. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 73–77.

Dasgupta, T.; Naskar, A.; Dey, L.; and Saha, R. 2018. Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, 93–102.

De Clercq, O.; Hoste, V.; Desmet, B.; Van Oosten, P.; De Cock, M.; and Macken, L. 2014. Using the crowd for readability prediction. *Natural Language Engineering* 20(3): 293–325.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*.

Dong, F.; Zhang, Y.; and Yang, J. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 153–162.

Farag, Y.; Yannakoudakis, H.; and Briscoe, T. 2018. Neural automated essay scoring and coherence modeling for adversarially crafted input. *arXiv preprint arXiv:1804.06898* .

Feng, L.; Jansche, M.; Huenerfauth, M.; and Elhadad, N. 2010. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd international conference on computational linguistics: Posters*, 276–284. Association for Computational Linguistics.

Ji, Y.; and Eisenstein, J. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 13–24.

Ji, Y.; and Smith, N. 2017. Neural discourse structure for text categorization. *arXiv preprint arXiv:1702.01829* .

Li, J.; and Hovy, E. 2014. A model of coherence based on distributed sentence representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2039–2048.

Liu, Y.; and Lapata, M. 2018. Learning structured text representations. *Transactions of the Association for Computational Linguistics* 6: 63–75.

Louis, A.; and Nenkova, A. 2012. A coherence model based on syntactic patterns. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1157–1168. Association for Computational Linguistics.

Marcu, D. 2000. *The theory and practice of discourse parsing and summarization*. MIT press.

Mesgar, M.; and Strube, M. 2018. A neural local coherence model for text quality assessment. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4328–4339.

Miltsakaki, E.; and Kukich, K. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering* 10(1): 25–55.

Moon, H. C.; Mohiuddin, M. T.; Joty, S.; and Xu, C. 2019. A Unified Neural Coherence Model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2262–2272.

Nadeem, F.; and Ostendorf, M. 2018. Estimating linguistic complexity for science texts. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 45–55.

Nguyen, D. T.; and Joty, S. 2017. A neural local coherence model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1320–1330.

Pappagari, R.; Zelasko, P.; Villalba, J.; Carmiel, Y.; and Dehak, N. 2019. Hierarchical Transformers for Long Document Classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 838–844. IEEE.

Petersen, C.; Lioma, C.; Simonsen, J. G.; and Larsen, B. 2015. Entropy and graph based modelling of document coherence using discourse entities: An application to IR. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, 191–200. ACM.

Socher, R.; Chen, D.; Manning, C. D.; and Ng, A. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*, 926–934.

Taboada, M.; and Mann, W. C. 2006. Applications of rhetorical structure theory. *Discourse studies* 8(4): 567–588.

Taghipour, K.; and Ng, H. T. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 1882–1891.

Tay, Y.; Phan, M. C.; Tuan, L. A.; and Hui, S. C. 2018. SkipFlow: incorporating neural coherence features for end-to-end automatic text scoring. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Todirascu, A.; François, T.; Bernhard, D.; Gala, N.; and Ligozat, A.-L. 2016. Are cohesive features relevant for text readability evaluation? In *26th International Conference on Computational Linguistics (COLING 2016)*, 987–997.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*.

Wang, Y.; Huang, S.; Li, G.; Deng, Q.; Liao, D.; Si, P.; Yang, Y.; and Xu, J. 2020. Cognitive Representation Learning of Self-Media Online Article Quality. *arXiv preprint arXiv:2008.05658* .

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1480–1489.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.