

# AN ENTITY-DRIVEN RECURSIVE NEURAL NETWORK MODEL FOR CHINESE DISCOURSE COHERENCE MODELING

Fan Xu, Shujing Du, Maoxi Li and Mingwen Wang

School of Computer Information Engineering, Jiangxi Normal University  
Nanchang 330022, China

## ABSTRACT

*Chinese discourse coherence modeling remains a challenge task in Natural Language Processing field. Existing approaches mostly focus on the need for feature engineering, which adopt the sophisticated features to capture the logic or syntactic or semantic relationships across sentences within a text. In this paper, we present an entity-driven recursive deep model for the Chinese discourse coherence evaluation based on current English discourse coherence neural network model. Specifically, to overcome the shortage of identifying the entity (nouns) overlap across sentences in the current model, our combined model successfully investigates the entities information into the recursive neural network framework. Evaluation results on both sentence ordering and machine translation coherence rating task show the effectiveness of the proposed model, which significantly outperforms the existing strong baseline.*

## KEYWORDS

*Entity, Recursive Neural Network, Chinese Discourse, Coherence*

## 1. INTRODUCTION

Discourse Coherence Modeling (DCM) aims to evaluate a degree of coherence among sentences within a discourse or text. It is considered one of the key problems in Natural Language Processing (NLP) due to its wide usage in many NLP applications, such as statistical machine translation<sup>[1]</sup>, discourse generation<sup>[2][3][4]</sup>, text automation summarization<sup>[3][5][6]</sup>, student essay scoring<sup>[7][8][9]</sup>.

In general, a coherent discourse generally has many similar components (lexical overlap or coreference) across sentences within a text, while incoherent discourse is the other one. Therefore, the traditional cohesion theory of Centering<sup>[10]</sup> driven and entity-based model<sup>[11][12][13][14]</sup> was proposed to capture the syntactic or semantic distribution of discourse entities (nouns) between two adjacent sentences in a text. Thereafter, many extension works were presented such as Feng and Hirst<sup>[15]</sup>'s multiple ranking model, Lin et al.<sup>[16]</sup>'s discourse relation-based approach, Louis and Nenkova<sup>[17]</sup>'s syntactic patterns-based model. However, the potential issue of the existing traditional coherence models need feature engineering, which is a time-consuming job.

In order to overcome the limitation of feature engineering issue, modern research tries to use neural network to extract the syntactic or semantic representation of a sentence automatically. Li et al.<sup>[18]</sup> proposed neural deep model to deal with English discourse coherence evaluation. However, their discourse coherence model only focuses on the distributed representation for

sentences, and did not consider the entity (nouns) distribution across sentences. In fact, the entities can be overlapped between two adjacent sentences, and are good insight to capture the coherence between adjacent two sentences as mentioned in traditional entity-based method. Therefore, we successfully integrate this kind of information into current recursive neural network framework. Evaluation results on both sentence ordering and machine translation coherence rating task show the effectiveness of the proposed model, which significantly outperforms the existing strong baseline.

Therefore, this paper tries to answer the following three questions:

- (1) Can the current English discourse coherence models (traditional or neural method) work for Chinese discourse coherence evaluation task?
- (2) Can the traditional entity based model be integrated into current deep model?
- (3) Which kind of word embedding works better for Chinese discourse coherence evaluation?

The rest of this paper is organized as follows. Section 2 reviews related work on discourse coherence modeling. Section 3 introduces the framework of our entity-driven recursive neural network based Chinese discourse coherence model. Section 4 describes the experiment results and detailed analysis. Finally, some conclusions are drawn in Section 5.

## 2. RELATED WORK

In this section, we describe the related work for discourse coherence modeling from traditional and neural network modes, respectively.

### 2.1. TRADITIONAL COHERENCE MODEL

The task of DCM was first introduced by Foltz et al.<sup>[19]</sup>. They formulated the discourse coherence as a function of semantic relatedness between two adjacent sentences within a text, and employed a vector-based representation of lexical meaning to compute the semantic relatedness. Since then, many supervised approaches to DCM, such as the entity-based model<sup>[11][12][13][15]</sup>, discourse relation-based model<sup>[16]</sup>, syntactic patterns-based model<sup>[17]</sup>, co reference resolution-based model<sup>[20][21]</sup>, content-based model via Hidden Markov Model (HMM)<sup>[3][22]</sup> and cohesion-driven based model<sup>[23]</sup> have been proposed in literature.

To be more specific, Barzilay and Lapata<sup>[11][12]</sup> presented an entity-based model to capture the distribution of discourse entities between two adjacent sentences within a text. As an extensive work of entity-based approach, Lin et al.<sup>[16]</sup> explored the function of discourse relations to revise the entity and to catch the behavior of discourse relation transfer among sentences. In addition, Feng and Hirst<sup>[15]</sup> showed that multiple ranking instead of pair wise ranking was effective for the DCM.

Differently, Louis and Nenkova<sup>[17]</sup> explored the function of syntactic structure in the DCM. Besides, Iida et al.<sup>[20]</sup> and Elsner et al.<sup>[21]</sup> demonstrated the importance of the usage of co reference resolution. In addition, Barzilay et al.<sup>[3]</sup> and Elsner et al.<sup>[22]</sup> showed that an Hidden Markov Model (HMM)-based content model can be used to capture the topic's transfer from the first sentence to the end sentence of a text, where topics were formulated as hidden states and sentences were treated as observations. Still, a potential issue of the HMM model is its domain-dependent mechanism. Also, Xu et al.<sup>[23]</sup> explored the impact of Halliday<sup>[24]</sup>'s Theme Structure Theory (TST) in English discourse coherence modeling. Their model shows the importance of the theme structure, a cohesion theory of Halliday's systemic-functional grammar, to DCM, and the appropriateness of theme and co reference based filtering mechanism.

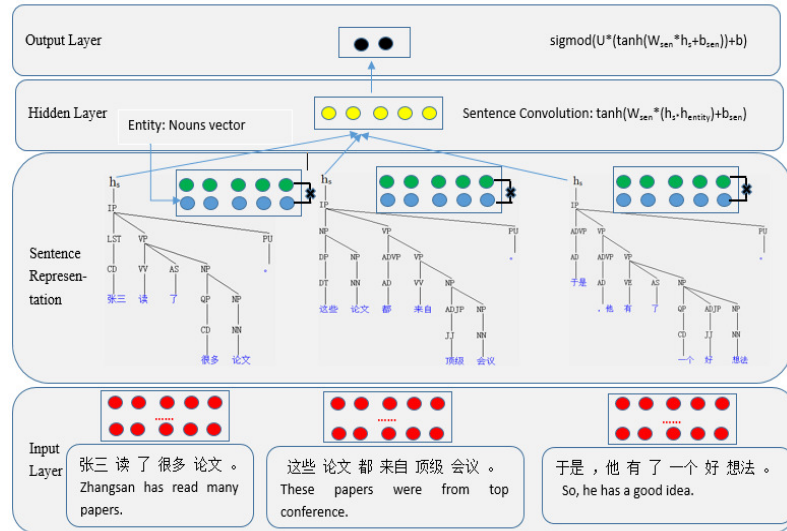


Figure 1: The framework of entity-driven recursive model for Chinese Discourse Coherence Modeling.

## 2.2. NEURAL COHERENCE MODEL

Recently, Li et al.<sup>[18]</sup> presented a neural deep model for English discourse coherence modeling. They demonstrated the effectiveness of both recurrent and recursive neural network (RNN) model for English situation.

However, as mentioned in the Section 1, their model did not consider the entity (nouns) distribution or entity overlap across sentences within a text. In fact, the entity overlap between two adjacent sentences indicates logical or semantic coherence for para text. Therefore, we successfully integrate these information into their model.

## 3. ENTITY-DRIVEN RNN COHERENCE MODEL

In this section, we describe our entity-driven RNN Chinese discourse coherence model.

### 3.1. FRAMEWORK

Figure 1 shows the entity-driven recursive deep model for Chinese discourse coherence modeling. Our deep model is based on Li et al.<sup>[18]</sup>'s English discourse coherence framework. On comparison, their model doesn't intensify the effectiveness of entities across each sentences in a text. Therefore, we successfully integrate the entities into current recursive neural network model.

### 3.2. SENTENCE REPRESENTATION

For the word-level representation, each word in a sentence can be represented by using a vector representation (or word embedding), and are able to capture the semantic meanings through toolkit, e.g. word2vec<sup>1</sup> or Glove<sup>2</sup>. More specifically, the word of a sentence can be represented using a specific vector embedding  $e_w = \{e_w^1, e_w^2, \dots, e_w^K\}$ , where  $K$  denotes the dimension of the word embedding.

<sup>1</sup><http://code.google.com/p/word2vec/>

<sup>2</sup><http://nlp.stanford.edu/projects/glove/>

For the sentence-level representation, as shown in Figure 1, the vector representation for the whole sentence is computed as a representation for each parent node based on its immediate children recursively in a bottom-up fashion until reaching the root of the tree. Concretely, for a given parent  $p$  in the tree and its two children  $c_1$  (associated with vector representation  $h_{c1}$ ) and  $c_2$  (associated with vector representation  $h_{c2}$ ), standard recursive network calculates  $h_p$  for  $p$  as follows:

$$h_p = f(W_{Recursive} \cdot [h_{c1}, h_{c2}] + b_{Recursive}) \quad (1)$$

where  $[h_{c1}, h_{c2}]$  refers to the concatenating vector for children vector  $h_{c1}$  and  $h_{c2}$ ;  $W_{Recursive}$  is a  $k \times 2K$  matrix and  $b_{Recursive}$  is the  $K \times 1$  bias vector;  $f(\cdot)$  is  $\tanh$  function.

### 3.3. ENTITY-DRIVEN SENTENCE CONVOLUTION

The framework treats a window of sentences as a clique  $C$  (sliding windows of  $L$  sentences) and associates each clique with a tag  $y_c$  that takes the value 1 if coherent, and 0 otherwise. As shown in Figure 1, each clique  $C$  takes as input a  $(L \times K) \times 1$  vector  $h_c$  by concatenating the embedding of all its contained sentences. The hidden layer takes as input  $h_c$  and performs the convolution using a non-linear  $\tanh$  function. The concatenating output vector for hidden layers, defined as  $q_c$ , can therefore be rewritten as:

$$q_c = f(W_{sen} \cdot (h_c \cdot h_{entity}) + b_{sen}) \quad (2)$$

where  $W_{sen}$  is a  $H \times (L \times K)$  dimensional matrix and  $b_{sen}$  is a  $H \times 1$  dimensional bias vector;  $H$  refers to the number of neurons in the hidden layer.

#### 3.3.1. ENTITY-DRIVEN MECHANISM

Firstly, we conduct vector summation operation for each nouns' word embedding to generate  $h_{entity}$  formulated as:

$$H_{entity} = e_{w_{NN1}} \oplus e_{w_{NN2}} \oplus \dots \oplus e_{w_{NNk}} \quad (3)$$

Then, we conduct element wise multiplication operation between  $h_c$  and  $h_{entity}$ .

The value of the output layer can be formulated as:

$$P(y_c = 1) = \text{sigmoid}(U^T q_c + b) \quad (4)$$

where  $U$  is an  $H \times 1$  vector and  $b$  denotes the bias;  $y_c$  with value 1 means the text is coherent, and 0 otherwise.

Therefore, the total coherence score for a given document is the probability that all cliques within the document are coherent, which is given by:

$$S_d = \prod_{C \in d} P(y_c = 1) \quad (5)$$

Finally, we can determine whether a text is coherent according to the value of their coherence score.

### 3.4. TRAINING AND OPTIMIZATION

The cost function for the model is given by:

$$J(\Theta) = \frac{1}{M} \sum_{c \in \text{trainset}} H_0 + \frac{Q}{2M} \sum_{\theta \in \Theta} \theta^2 \quad (6)$$

$$H_0 = -y_c \log[p(y_c = 1)] - (1 - y_c) \log[1 - p(y_c = 1)] \quad (7)$$

where  $\Theta = [W_{\text{Recursive}}, W_{\text{sen}}, U_{\text{sen}}]$ ;  $M$  denotes the number of training samples.

We adopt the widely applied optimization diagonal variant of AdaGrad (Duchi et al.<sup>[25]</sup>) to optimize the loss function.

## 4. EXPERIMENTS

In this section, we demonstrate the effectiveness of our discourse coherence model through both sentence ordering and machine translation coherence rating tasks. The former aims to discern an original text from a permuted ordering of its sentences, while the latter aims to discern a human or reference translation from automatically machine generated translation.

### 4.1. DATASET

**Sentence Ordering Dataset:** We select documents for Chinese Treebank 6.0 from Linguistic Data Consortium (LDC) with catalog number LDC2007T36 and ISBN1-58563-450-6. We select the 100 documents from chtb\_2946 to chtb\_3045 as our training dataset, and 100 documents from chtb\_3046 to chtb\_3145 as our testing dataset. The sentences in each source file will be permuted at most 20 times. The total number of testing texts is 1027. The average number of sentence are 10.33 and 13.56 for training set and testing set, respectively. In the evaluation, we

consider the original texts are more coherent (positive instances) than the permuted ones (negative instances).

**Machine Translation Dataset:** Similarly, we extract documents for NIST Open Machine Translation 2008 Evaluation (MT08) Selected Reference and System Translations from Linguistic Data Consortium (LDC) with catalog number LDC2010T01 and ISBN1-58563-533-2. Therein, the English-to-Chinese language pairs have 127 documents with 1830 segments, output from 11 machine translation systems. The average number of sentence are 13.38 and 13.39 for training set and testing set, respectively. In evaluation, we consider the human or reference translation texts are more coherent than the machine generated one.

### 4.2. EXPERIMENTAL SETTINGS

**Initialization:** Similar to Li et al.<sup>[18]</sup>, the parameter  $W_{\text{sen}}$ ,  $W_{\text{recursive}}$  and  $h_0$  are initialized by randomly drawing from the uniform distribution. The number of hidden layer  $H$  is set to 100. Learning rate in the optimization process is set to 0.01, and batch size is set to 20. Differently, word embedding  $\{e\}$  for Chinese are trained using word2vec and Glove respectively. The dimension for word embedding is 50 or 100. The window size  $L$  is 3 or 5.

**Evaluation Metric:** We report system's performance using accuracy, which is the ratio of the number of the selected original text/translation document divided by the total number of texts/translation document.

**Baseline System 1:** Entity graph based model<sup>[14]</sup> which has been demonstrated as a simple but effective implementation of the entity-based coherence model. We re-implement their method in this paper based on publicly available code<sup>3</sup>.

**Baseline System 2:** Another baseline, Li et al.<sup>[18]</sup>'s recursive neural model, which did not consider the entity transition information. We transplant their English discourse coherence framework to Chinese situation. Furthermore, we successfully integrate the entity information into their deep model.

In addition, we employ Stanford parser<sup>4</sup> to generate sentence-level constituent parser tree and generate the part-of-speech to get the entities (nouns) occur in each sentence, and use utility ICTCLAS<sup>5</sup> to conduct Chinese word segmentation.

### 4.3. EXPERIMENT RESULTS

In this section, we report the experiment results for the Chinese discourse coherence modeling on both sentence ordering and machine translation coherence rating task.

#### 4.3.1. RESULTS ON SENTENCE ORDERING

Table 1 shows the performance of our entity-driven deep model using different windows size, different dimension, and with different type of word embedding.

Table 1: The performance under different settings on sentence ordering task.

	dimension=50		dimension=100	
	Window size			
	3	5	3	5
Glove	56.03	49.37	<b>65.67</b>	52.47
word2vec	57.52	48.68	65.56	53.50

As shown in Table 1, it shows that:

#### (1) Dimension

Generally speaking, the performance increases with the increment of the dimension. In fact, the larger the dimension, the more representative ability it is.

#### (2) Window size

The performance decreases with the increment of the window size, and the best performance yields at the window size with 3. It is mostly caused by the local entity distribution characteristic demonstrated by Barzilay and Lapata<sup>[11][12]</sup>, Guinaudeau and Strube<sup>[14]</sup>. As the increment of the number of the window size, the entity co-occurrence decreases accordingly.

<sup>3</sup><http://github.com/karins/CoherenceFramework>.

<sup>4</sup><http://nlp.stanford.edu/software/lex-parser.shtml>.

<sup>5</sup><http://ictclas.nlpir.org/downloads>

Table 2, below, lists the performance comparison among our model and the current baseline model (traditional model and neural network model).

Table 2: Performance comparison among different coherence model on sentence ordering task; Performance that is significantly superior to baseline systems ( $p < 0.05$ , using paired t-test for significance) is denoted by \*.

Entity graph based model <sup>[14]</sup>	67.78
Li et al. <sup>[13]</sup> (Glove word embedding)	65.67
Li et al. <sup>[13]</sup> (word2vec word embedding)	65.56
Our combined model	<b>67.16*(Li's model)</b>

It shows that our combined model significantly outperforms the current deep model for Chinese discourse coherence modeling, which demonstrates the effectiveness and importance of the entity distribution across sentences. Interestingly, the traditional entity based model also works for Chinese discourse coherence evaluation, which doesn't work fine for English situation. This is mostly caused by the entity distribution are obvious in Chinese discourse than in English text.

#### 4.3.2. RESULTS ON MACHINE TRANSLATION COHERENCE RATING

Table 3, below, lists the performance of our model and the baseline model.

Table 3: Performance comparison among different coherence model on machine translation coherence rating task with dimension equals to 100; Performance that is significantly superior to baseline systems ( $p < 0.05$ , using paired t-test for significance) is denoted by \*.

Entity graph based model <sup>[14]</sup>	68.50
Li et al. <sup>[13]</sup> (Glove word embedding)	70.08
Li et al. <sup>[13]</sup> (word2vec word embedding)	68.50
Our combined model	<b>72.44*</b>

In fact, discourse coherence evaluation for the machine translation task is more common than the sentence ordering task evaluation. As the results show in Table 3, again, our model significantly outperforms the current model. Also, our model significantly outperforms the traditional entity-based model. It is mostly caused by the entity distribution is not obvious in the text generated by the machine. But the entity (nouns) information still can be integrated into current recursive neural network model.

## 5. CONCLUSIONS AND FUTURE WORK

In this paper, we present an entity-driven recursive deep model for Chinese discourse coherence modeling. We successfully integrate the entities across each sentence into current recursive neural framework. Evaluation results on both sentence ordering and machine translation coherence rating task show the effectiveness of the proposed model. Our future work is to integrate the co reference mechanism into current combined recursive neural network model, together with other coherence evaluation task.

## ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China under Grant No.61402208, No.61462045, No.61462044, No.61662030, the Natural Science Foundation and Education Department of Jiangxi Province under Grant No. 20151BAB207027 and GJJ150351, and the Research Project of State Language Commission under Grant No.YB125-99.

## REFERENCES

- [1] Heidi J. Fox,(2002),“Phrasal Cohesion and Statistical Machine Translation”, In Proceedings of EMNLP, pages304-311.
- [2] Radu Soricut and Daniel Marcu,(2006),“Discourse Generation Using Utility-Trained Coherence Models”, In Proceedings of COLING-ACL, pages803-810.
- [3] Regina Barzilay and Lillian Lee,(2004),“Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization”, In Proceedings of NAACL-HLT, pages113-120.
- [4] Jiwei Li, Minh-Thang Luong and Dan Jurafsky,(2015),“A Hierarchical Neural Autoencoder for Paragraphs and Documents”, In Proceedings of ACL, pages1106-1115.
- [5] Zi-Heng Lin, Hwee Tou Ng and Min-Yen Kan,(2012),“Combining Coherence Models and Machine Translation Evaluation Metrics for Summarization Evaluation”, In Proceedings of ACL,pages 1006–1014.
- [6] Danushka Bollegala, Naoaki Okazaki and Mitsuru Ishizuka,(2006),“A Bottom-Up Approach to Sentence Ordering for Multi-Document Summarization”,In Proceedings of ICCL-ACL,pages 385-392.
- [7] Helen Yannakoudakis and Ted Briscoe,(2012),“Modeling coherence in ESOL learner texts”, In Proceedings of ACL, pages33-43.
- [8] Jill Burstein,Joel Tetreault and Slava Andreyev,(2010),“Using Entity-Based Features to Model Coherence in Student Essays”, In Proceedings of NAACL-HLT, pages681-684.
- [9] Derrick Higgins, Jill Burstin,Daniel Marcu and Claudia Gentile,(2004),“Evaluating Multiple Aspects of Coherence in Student Essays”, In Proceedings of NAACL-HLT,pages185-192.
- [10] Barbara J. Grosz, Scott Weinstein and Aravind K. Joshi,(1995),“Centering:A Framework for Modeling the Local Coherence of Discourse”,Computational Linguistics, 21(2):203-225.
- [11] Regina Barzilay and Mirella Lapata,(2005),“Modeling Local Coherence: An Entity-Based Approach”, In Proceedings of ACL,pages 141-148.
- [12] Regina Barzilay and Mirella Lapata,(2008),“Modeling Local Coherence: An Entity-Based Approach”,Computational Linguistics, 34(1):1-34.
- [13] Mirella Lapata and Regina Barzilay,(2005), “Automatic Evaluation of Text Coherence: Models and Representations”, In Proceedings of IJCAI, pages 1085-1090.
- [14] Camille Guinaudeau and Michael Strube, (2013), “Graph-based local coherence modeling”, In Proceedings of ACL, pages 93–103.
- [15] Vanessa Wei Feng and Graeme Hirst,(2012),“Extending the Entity-based Coherence Model with Multiple Ranks”, In Proceedings of EACL, pages 315-324.
- [16] Zi-Heng Lin, Hwee Tou Ng and Min-Yen Kan,(2011),“Automatically Evaluating Text Coherence Using Discourse Relations”, In Proceedings of ACL, pages 997-1006.
- [17] Annie Louis and Ani Nenkova,(2012),“A coherence model based on syntactic patterns”, In Proceedings of EMNLP-CNLL, pages1157-1168.
- [18] Jiwei Li and Eduard Hovy,(2014),“A Model of Coherence Based on Distributed Sentence Representation”, In Proceedings of EMNLP, pages2039-2048.
- [19] Peter W. Foltz, Walter Kintsch and Thomas K. Landauer,(1998),“The measurement of textual coherence with latent semantic analysis”,Discourse Processes,25(2&3): 285-307.
- [20] Ryu Iida and Takenobu Tokunaga,(2012),“A Metric for Evaluating Discourse Coherence based on Coreference Resolution”, In Proceedings of COLING, Poster, pages483-494.
- [21] Micha Elsner and Eugena Charniak,(2008),“Coreference-inspired Coherence Modeling”, In Proceedings of ACL 2008, Short Papers, pages41-44.
- [22] Micha Elsner, Joseph Austerweil and Eugene Charniak,(2007),“A Unified Local and Global Model for Discourse Coherence”, In Proceedings of NAACL, pages436-443.



- [23] Fan Xu, Qiaoming Zhu, Guodong Zhou and Mingwen Wang,(2014),“Cohesion-driven Discourse Coherence Modeling”, Journal of Chinese Information Processing, 28(3):11-21.
- [24] M. A. K. Halliday,(1994),“An Introduction to Functional Grammar”, Hodder Education Press, London, United Kingdom.
- [25] John Duchi, Elad Hazan and Yoram Singer,(2011),“Adaptive subgradient methods for online learning and stochastic optimization”,The Journal of Machine Learning Research,12:2121-2159.

## AUTHORS

**Fan Xu** holds a Doctoral Degree (Ph.D.) in Computer Science from Soochow University, China. His areas of research interest includes Natural Language Processing, Chinese Information Processing, Discourse Analysis, and Speech Recognition. At present he is working as Lector, School of Computer Information Engineering, Jiangxi Normal University, China. He is member of various professional bodies including ACL, IEEE, and ACIS.



**Shujing Duis** a Master of Computer Science of Jiangxi Normal University, China. Her research interest includes Natural Language Processing, Chinese Information Processing, Discourse Analysis



**Maoxi Li** holds a Doctoral Degree (Ph.D.) in Computer Science from Chinese Academy of Sciences. His areas of research interest includes Machine Translation and Natural Language Processing. At present he is working as Associate Professor, School of Computer Information Engineering, Jiangxi Normal University, China.



**Mingwen Wang** holds a Doctoral Degree (Ph.D.) in Computer Science from Shanghai Jiaotong University, China. His areas of research interest includes Machine Learning, Information Retrieval, Natural Language Processing, Image Processing, and Chinese Information Processing. At present he is working as Professor, School of Computer Information Engineering, Jiangxi Normal University, China. He is member of various professional bodies including ACL, IEEE, CCF, and ACIS.

