A Maximum Entropy Approach to Discourse Coherence Modeling

Rui Lin[1(✉)], Muyun Yang[1], Shujie Liu[2], Sheng Li[1], and Tiejun Zhao[1]

[1] Machine Intelligence & Translation Lab,
Harbin Institute of Technology, Harbin 150001, China
{linrui,ymy}@mtlab.hit.edu.cn, {lisheng,tjzhao}@hit.edu.cn
[2] Microsoft Research, Harbin, China
shujliu@microsoft.com

Abstract. This paper introduces a maximum entropy method to Discourse Coherence Modeling (DCM). Different from the state-of-art supervised entity-grid model and unsupervised cohesion-driven model, the model we proposed only takes as input lexicon features, which increases the training speed and decoding speed significantly. We conduct an evaluation on two publicly available benchmark data sets via sentence ordering tasks, and the results confirm the effectiveness of our maximum entropy based approach in DCM.

Keywords: Discourse coherence · Maximum entropy · Sentence ordering

1 Introduction

A high performance discourse coherence model (DCM) is important for natural language processing and generating tasks of multi-sentence document. Coherence, both logically and syntactically, makes a text meaningful [17]. For a well-written text, if we keep the word order in each sentence and swap the sentences randomly, the original text could be totally unreadable. The following two examples will show what coherence is:

- Example 1: You want my name? My name is John Smith.
- Example 2: You want my name? I am 24 years old now.

Each of the two simple texts contains only two sentences. Considering each of the four sentences above, we can see that all of them are correct both logically and syntactically. But regarding the two adjacency sentence as a whole text, example 1 is coherent and easy to understand while the example 2 makes people confused and unable to get the point. Therefore, coherence is essential in generating readable text with reasonable sentence order. The discourse coherence is a key requirement for text generating system so that it is widely used in natural language processing and generating applications such as: statistical machine translation [1], discourse generation [2] and summary [3].

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 3–11, 2015.
DOI: 10.1007/978-3-319-25207-0_1

Owing to the importance of discourse coherence modeling, a variety of coherence theories have been developed since 1980s. Halliday and Hasan aruge that text is not consists of irrelevant sentences in 1980: each sentence plays with important role regards to the whole[4]. Beaugrande and Dressle[5] point out 7 basic features of a coherent discourse in 1981, which are :cohesion, coherence, intentionality, acceptability, informativity, situationality, and intertextuality, in which the cohesion and coherence are meaningful in natural language processing. A quite influential theory, Rhetorical Structure Theory (RST)[6], is proposed by Mann and Thompson. RST defines 25 relations that govern clause interdependencies and ordering. These relations can be represent as a tree structure. Cristea and Romary propose Veins Theory which is also based on RST[7]. Another influential theory is Centering Theory (CT)[8] proposed by Grosz et al. CT use entity to capture the coherence between sentences in a document. Besides these method, there are also many others such as Dscourse Representation Theory (DRT) proposed by Kamp[9].

Based on these theories, some computable approaches have been developed. Barilay and Lapata[10] propose an entity-grid method which is a recent popular method based on CT to do discourse coherence modeling. Their method capture the information on the role the entity plays to judge whether a document is coherence or not. Many following efforts adopt the same framework and extend the entity-grid method by adding useful features such as discourse relations[11], multiple rank[12], named entities[13] and graph model[14]. Louis and Nenkova do coherence modeling using Hidden Markov Model (HMM) with syntactic features[15] . Xu et al.[16] proposed an unsupervised cohesion-driven method. It can get a satisfying performance when coreference resolution. Li and Hovy propose a recurrent and recursive neural network coherence model which improves the accuracy significantly[17].

All these method mentioned above need a high cost preprocess such as syntax parsing and coreference resolution which makes the training inefficient. In the other hand, most of these methods regard the whole document as inseparable element which can't capture the inner connection with high performance. And it is not suitable for statistical machine translation or text generating for there is no completed sentence or discourse for the decoding period in these tasks.

In this paper, we propose a maximum entropy based discourse coherence modeling method. Our maximum entropy based method modeling the discourse coherence with lexicon features instead of extract features from the whole document. In contrast to the previous methods, our approach can train the discourse coherence model without preprocess such as syntax parsing and coreference resolution. In this sense, out method is language independent since it uses only lexicon feature. Compared with the state-of-the-art methods in the same benchmarks, our approach also performs a better result as well as a faster training speed.

The rest of this paper is organized as follows: Section 2 introduces work related to doing discourse coherence modeling. Section 3 introduces how we do discourse coherent modeling with maximum entropy model. Section 5 presents our experiments and the results. Finally, we conclude in Section 7.

2 Related Work

Early works about discourse coherence model are described by linguists, such as RST, CT and DRT. These works only illustrate some concepts of discourse modeling and few of them are computable methods.

A recent popular approach is entity-grid method proposed by Barzilay and Lapata [10]. Their method represent a document with an entity-grid. An entity-grid is a table that each row of it denotes a sentence and each column of it denotes an entity. The element in the entity-grid has four states: O stands for object; S stands for Subject; X stands for neither object nor subject and – stands for absent. They get a discourse feature vector by counting the state transition frequency. A support vector machine (SVM) is used to judge whether a document is coherent or not. This approach can add many other features such as discourse relations[11], multiple rank[12], named entities[13] and graph model[14]. But it need high cost preprocess and is not suitable for all tasks.

Louis and Nenkova[15] propose a HMM based coherence model which is different from entity-grid methods. They use syntax features which convert the sentences in the document into production type. By clustering these productions they get some classed and regard them as hidden state. The productions are regarded as observation and the document is regard as a sequence data. They train a HMM to compute the coherent. Their approach only uses syntax features but ignore the semantics features.

Besides the supervised approaches mentioned above, an unsupervised method, cohesion driven approach, is proposed by Xu et al.[16]. Their method divides a sentence into two parts, theme and rheme. The coherent score of adjacent sentences is computed via thematic progression. The coherent of the whole document is computed with each score of adjacent sentences. This method can easily applied but the accuracy is lower than the supervised approaches and it also need high cost preprocess.

To further improve the performance of the discourse coherence model, Li and Hovy[17] propose a neural network coherence model. Their method examines a recurrent and a recursive neural network to train sentence embedding to represent a sentence. And a neural network classifier takes a slide window of these embeddings as input to compute the coherent probability of the window of sentences. After sliding all the sentences in the document, the coherent can be computed from the score of each window. This approach has the state-of-the-art performance but it is a deep learning method so the training and decoding speed is quite slow. It is a method with high computational complexity.

3 Maximum Entropy Based Discourse Coherence Model

Beaugrande and Dressle[5] point out 7 basic features of a coherent discourse, in which, cohesion and coherence explains the relationship between sentences in the same document. Therefore, in a coherence document, the words in the current sentence are chosen depending on the previous sentences.

3.1 Discourse Model

To model the coherence of sentences in the document D, which contains sentences $S_1, S_2, \ldots, S_n$, we need to maximize the objective function as fellow:

$$p(D) = p(S_1, S_2, \ldots, S_n) = p(S_1) \cdot p(S_2|S_1) \cdot \ldots \cdot p(S_n|S_1, S_2, \ldots, S_{n-1}) \quad (1)$$

Where $p(D)$ denotes the probability of the coherence of document D. From the Eq.(1) we can find that this objective function is almost the same as the objective function of language modeling. The difference is that we compute the probability of the document here while language model computes the probability of sentence. It is too complex to compute the probability directly with Eq.(1). So we limit the history length as language model does. After introduced this feature, the equation can be simplified as follow:

$$p(D) \approx p(S_1) \cdot p(S_2|S_1) \cdot p(S_3|S_2) \cdot \ldots \cdot p(S_n|S_{n-1}) = \prod_{k=1}^{n} p(S_k|h) \quad (2)$$

Where h denotes the history. Like language modeling, the longer the history length is, the closer to original objective function this function will be. But considering the sparsity of the sentence, the longer history will make the model over-fitting. So our model uses bigram history. To improve the performance of the model, the sentence also should be simplified as a vector. Considering the computational complexity, we use bag of words to represent a sentence in our model. So the Eq.(2) can be simplified as follow:

$$p(D) \approx \prod_{k=1}^{n} p(BoWS_k|BoWh) \quad (3)$$

where $BoWS_k$ denotes the bag of words of k-th sentence. To maximize the $p(D)$, we can maximize each multiplier separately. And bag of words can be easily converted into a feature vector. Here we introduce the maximum entropy model to model the discourse.

3.2 Maximum Entropy Based Discourse Coherence Model

Ep.(2) shows the similarity between discourse coherence model and language model. In this paper, we introduce the language model approach to capture the discourse coherence. Ep.(3) shows the bag of words representation of sentences, conventional n-gram language model is not suitable for this. Maximum entropy language model can capture more information and we can add any features to maximum entropy model[18]. So we decide to introduce maximum entropy model to doing discourse coherence modeling.

For a maximum entropy language model, the probability of current word w given history h is computed as follow:

$$p(w|h) = {}^{1}/_{Z(h)} \cdot \exp\left(\sum_i \lambda_i f_i(h, w)\right) \quad (4)$$

where $Z(h)$ denotes the normalization factor of history, f_i denotes the i-th feature functions. Many different features can be added into maximum entropy language model, both n-gram features and long distance trigger. The more features added, the more information will be captured. We can get a good performance with an appropriate feature set. Combining Ep.(2), Ep.(3) and Ep.(4), we can compute the conditional probability of sentence as follow:

$$p(S_k|S_{k-1}) = {}^{1}/_{Z(S_{k-1})} \cdot \exp(\sum_i \lambda_i f_i(S_{k-1}, S_k))$$
$$= {}^{1}/_{Z(BoWS_{k-1})} \cdot \exp(\sum_i \lambda_i f_i(BoWS_{k-1}, BoWS_k)) \quad (5)$$

where $p(S_k|S_{k-1})$ denotes the probability of coherence while current sentence is S_k given history S_{k-1}, different from maximum entropy language model. Because of the difficultness representation of sentence and the data sparsity, we use bag of words to represent a sentence.

To better capture the discourse information, we introduce two feature functions $f_{w_k}(BoWS_{k-1}, BoWS_k)$ and $f_{w_{k-1}}(BoWS_{k-1}, BoWS_k)$ as follow:

$$f_{w_k}(BoWS_{k-1}, BoWS_k) = \begin{cases} 1 & w_k \in BoWS_k \\ 0 & w_k \notin BoWS_k \end{cases} \quad (6)$$

$$f_{w_{k-1}}(BoWS_{k-1}, BoWS_k) = \begin{cases} 1 & w_{k-1} \in BoWS_{k-1} \\ 0 & w_{k-1} \notin BoWS_{k-1} \end{cases} \quad (7)$$

Ep.(6) captures the information of current sentence while Ep.(7) captures the information of history. We can model the discourse coherence with this two feature functions well. We can get the coherent score of adjacent sentences after training the model. And the coherent score of the whole document can be computed with each adjacent score multiplied.

3.3 Model Training

We train our model with an open source maximum entropy tool, Maxent [文献或者链接]. The positive examples are sampled from the sentence pairs of the original documents while the negative examples are sampled from the sentence pairs of the permutation randomly. The ratio of positive example to negative examples is about 1. The training algorithm is default as L-BFGS, maximum number of iteration is 300.

4 Experiment

We conduct a sentence ordering task with two different corpora to evaluate our model. Sentence ordering is to find the original ordered document from a pair of articles. An article pair consists of one original document order and a random permutation of the sentences from the same document. Our approach is predicated on the assumption that the original article is always more coherent than a random permutation, which

has been verified in Lin et al.'s work[11]. We use the accuracy to evaluate the performance of sentence ordering task. Accuracy defines as the ratio of the correct number of pair to the total number of pair.

4.1 Dataset

Following the former experimental settings[10-17], the two different corpora we use for evaluation is from the Barzilay and Lapata[10] which contains original documents and generated permutation of the documents. One corpora contain reports about earthquake from the Associated Press and the other contains reports on airplane accidents from the National Transportation Safety Board. Each document contains about 10 sentences with clear structure. The information of the dataset are shown in the below table.

Table 1. Dataset Information

	Original	Permutation
Earthquake Train	100	2035
Earthquake Test	99	1956
Accident Train	100	2100
Accident Test	100	1986

From the Table 1 we can see that there is a total of 2135 training documents and 1956 test pairs of articles for earthquake corpare. For accdent corpora, there is a total of 2200 training documents and 1986 test pairs of articles.

4.2 Model Comparison

Table 2 shows the performance of our approach and other related work, include:

Recurrent and Recursive Neural Network Coherence Model: Li and Hovy[17] propose a neural network coherence model which obtains the best performance. They use a recurrent or recursive neural network to convert a sentence to a tree structure with word embedding. And the root node is a vector of a sentence. Another neural network classifier is applied to generate the probability of the coherence of the slide window. Comparing to former approaches, they model can be learned without feature engineering. The results are taken directly from Li and Hovy's paper[17].

Entity-Grid Model: This model is proposed by Barzilay and Lapata[10] in 2005. Only the neural network model gets a better performance than this method considering the average accuracy. This approach obtains good performance when coreference resolution, expressive syntactic information and salience-based features are incorporated. They use the n-gram transition as the feature vector and apply a SVM ranker to judge which document is better. The results are taken directly from Barzilay and Lapata's paper[10].

HMM: Hidden Markov Coherence Model proposed by Louis and Nenkova[15] capture the hidden state transition probability in the coherent context using syntactic features. They use a production to represent a sentence. And the productions are clustered as the hidden state. The results are taken directly from Louis and Nenkova's paper[15].

Table 2. Sentence ordering task experimental result

	Accident	Earthquake	Average
Entiy-Grid[10]	**0.904**	0.872	0.888
HMM[15]	0.822	0.938	0.880
Cohesion-driven[16]	0.886	0.848	0.867
Recursive[17]	0.864	**0.976**	0.920
Recurrent[17]	0.840	0.951	0.895
ME model	0.877	0.973	**0.925**
CIME model	0.870	0.970	0.920
ME + Entity-Grid	0.877	0.973	**0.925**

Cohesion-Driven Model: Xu et al.[16] propose a Cohesion-driven discourse coherence model. They divide a sentence into two parts, theme and rheme. The coherent score of adjacent sentences is the cosine similarity computed by thematic progression. The coherent of the whole document is computed with each score of adjacent sentences. The performance will be increased with coreference resolution applied. The results are taken directly from Xu et al.'s paper[16].

As can be seen in Table 1, our maximum entropy based approach outperforms all existing baselines and obtains a state-of-art performance. Our method's accuracy is a little low than the neural network model in earthquake corpora and gets a better performance in average.

Comparing to other baselines, our method can process the data without any cost preprocess such as syntax parsing and coreference resolution. So our model can be added into any other natural language processing application easily. And the training and decoding speed is much quicker than other methods.

Comparing to the recurrent and recursive neural network method, the maximum entropy based approach we proposed has little training and decoding cost. The training and decoding speed is 30x~60x faster than the recurrent and recursive neural network method. While the maximum entropy model is similar to two layer neural network. So our model can be easily extended to a deep learning model.

We also train a case insensitive maximum entropy (CIME) model with lowercased corpora. Comparing to the original case sensitive maximum entropy model we can find that the case sensitive model outperforms the case insensitive model a little bit. This is because the uppercase name entity plays an important role in discourse coherence modeling so it should be treated as a feature.

To better evaluate our model, we also extend the entity-grid method with our model. For the entity-grid method, it generates a feature vector for each document. We compute the coherence score for each document with our model and integrate this score into the feature vector of entity-grid method. We can find that after adding our features, the accuracy of entity-grid method stays the same as our maximum entropy model.

It means that our maximum entropy model has the strong features that makes the model has an extremely high accuracy on training data. The maximum entropy model is more suitable for the sentence ordering tasks than entity-grid model.

5 Conclusion

In this paper, we compare the existing discourse coherence modeling methods and the discourse coherence application. We conduct a maximum entropy based discourse coherence model without cost preprocessing such as coreference resolution or syntax parsing. Experiment with sentence ordering task, our model can get a good performance with only lexicon features applied in two different corpora. The training and decoding for our model are also efficient.

In the future, we try to apply our maximum entropy based discourse coherence model to statistical machine translation or discourse generation, hope to get a better performance. Also we decide to convert the maximum entropy model to two-layer neural network model and adding other neural network techniques to improve our model's performance.

Acknowledgments. Thanks for the three anonymous reviewers for their efforts. This paper is supported by the project of National Natural Science Foundation of China (Grant No. 61272384, 61370170 &61402134).

References

1. Tu, M., Zhou, Y., Zong, C.: Enhancing grammatical cohesion: generating transitional expressions for SMT. In: 52nd Annual Meeting of the ACL, Baltimore, USA (2014)
2. Prasad, R., Bunt, H.: Semantic relations in discourse: the current state of ISO 24617-8. In: Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11), pp. 80–92 (2015)
3. Lin, Z.H., Liu, C., Ng, H.W., Kan, M.Y.: Combining coherence models and machine translation evaluation metrics for summarization evaluation. In: Proceedings of the ACL. Association for Computational Linguistics, Jeju, pp. 1006–1014 (2012)
4. Halliday, M.A.K., Hasan, R.: Text and context: aspects of language in a social-semiotic perspective. Sophia Linguistica **6**, 4–91 (1980). Working Papers in Linguistics Tokyo
5. De Beaugrande, R.A., Dressler, W.U.: Introduction to text linguistics. Longman, London (1981)
6. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: Toward a functional theory of text organization. Text **8**(3), 243–281 (1988)
7. Cristea, D., Ide, N., Romary, L.: Veins theory: A model of global discourse cohesion and coherence. In: Proceedings of the 17th international conference on Computational linguistics. Association for Computational Linguistics, vol. 1, pp. 281–285 (1998)
8. Grosz, B.J., Joshi, A.K., Weinstein, S., et al.: Centering: A Framework for Modelling the Local Coherence of Discourse. Computational Linguistics **21**(2), 203–225 (1995)
9. Kamp, H., Kamp, H.: Discourse Representation Theory: What it is and Where it Ought to Go. Natural Language at the Computer **320**(1), 84–111 (1988)

10. Barzilay, R., Lapata, M.: Modeling local coherence: an entity-based approach. Computational Linguistics **34**(1), 1–34 (2008)
11. Lin, Z.H., Ng, H.T., Kan, M.Y.: Automatically evaluating text coherence using discourse relations. In: Proceedings of the ACL. Association for Computational Linguistics, Portland, pp. 997–1006 (2011)
12. Feng, V.W., Hirst, G.: Extending the entity-based coherence model with multiple ranks. In: Proceedings of the EACL. Association for Computational Linguistics, Avignon, pp. 315–324 (2012)
13. Eisner, M., Charniak, E.: Extending the entity grid with entity-specific features. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, short papers, vol. 2. Association for Computational Linguistics, pp. 125–129 (2011)
14. Guinaudeau, C., Strube, M.: Graph-based Local Coherence Modeling. In: ACL, vol. 1, pp. 93–103 (2013)
15. Louis, A., Nenkova, A.: A coherence model based on syntactic patterns. In: Proceedings of the EMNLP-CNLL. Association for Computational Linguistics, Jeju, pp. 1157–1168 (2012)
16. Xu, F., Zhu, Q., Zhou, G., et al.: Cohesion-driven Discourse Coherence Modeling. Journal of Chinese Information **28**(3), (2014)
17. Li, J., Hovy, E.: A model of coherence based on distributed sentence representation. In: Proceedings of the EMNLP (2014)
18. Rosenfeld, R.: A maximum entropy approach to adaptive statistical language modelling. Computer Speech & Language **10**(3), 187–228 (1996)