Discourse Element Identification in Student Essays based on Global and Local Cohesion

Wei Song[†], **Ruiji Fu**[‡], **Lizhen Liu**[†], **Ting Liu**[§]
[†]Information Engineering, Capital Normal University, Bejing 100048, China
[‡]Iflytek Research Beijing, Beijing 100083, China
[§]Harbin Institute of Technology, Harbin 150001, China
{wsong, lzliu}@cnu.edu.cn, rjfu@iflytek.com, tliu@ir.hit.edu.cn

Abstract

We present a method of using cohesion to improve discourse element identification for sentences in student essays. New features for each sentence are derived by considering its relations to global and local cohesion, which are created by means of cohesive resources and subtopic coverage. In our experiments, we obtain significant improvements on identifying all discourse elements, especially of +5% $F1$ score on *thesis* and *main idea*. The analysis shows that global cohesion can better capture *thesis statements*.

Element	Definition
Introduction (I)	introduces the background and/or grabs readers' attention
Prompt (P)	restates or summarize the prompt
Thesis (T)	states the author's main claim on the issue for which he/she is arguing
Main idea (M)	asserts foundational ideas or aspects that are related to the thesis
Supporting idea (S)	provides evidence to explain or support the thesis and main ideas
Conclusion (C)	concludes the whole essay or one of the main ideas
Other (O)	doesn't fit into the above elements or makes no meaningful contribution

Table 1: Definitions of discourse elements.

1 Introduction

Automatic discourse analysis of student essays can benefit many downstream applications such as essay rating, text organization assessment and writing instruction. In this paper we focus on identifying discourse elements for sentences in persuasive essays written by Chinese high school students. **Discourse elements** represent the contributions that sentences can make to text organization. Typical discourse elements and their functions in persuasive writing are summarized in Table 1.

Previous work mainly exploits the properties of a sentence itself or adjacent sentences for this task. In this work, we explore cohesion to express relations among sentences through the whole text. Cohesion can be defined as a set of resources linking within a text that organize the text together (Halliday and Hasan, 1976). It can be achieved through the use of reference, ellipsis, substitution, conjunction and lexical cohesion. Among them, lexical cohesion has been widely used for modeling local coherence and applied to related applications (Barzilay and Elhadad, 1999; Barzilay and Lapata, 2008; Galley et al., 2003; Hsueh et al., 2006; Filippova and Strube, 2006). Since cohesion is closely related to the structure of text (Morris and Hirst, 1991), it motivates us to explore similar techniques for discourse element identification. In addition, its ease of implementation is also attractive. Other options for representing text structure such as full-text discourse parsers (Marcu, 2000) may be not available or don't have satisfied performance, especially for non-English languages.

However, modeling local coherence alone is not adequate to distinguish discourse elements in persuasive essays. For example, a *main idea* may be followed by a *supporting idea* sentence. The two sentences can be coherent but their discourse elements are different. To deal with this, global cohesion should be exploited. Considering that in persuasive writing, *thesis*, *main ideas* and *conclusion*, which are termed *thesis statements* by Burstein et al. (2001), are expected to relate to each other (Higgins et al., 2004). It is likely that cohesive relations exist among them through the whole text.

We make a focused contribution by investigating global and local cohesive relations. We create sentence chains based on cohesive resources and examine whether the chains represent global cohesion or local cohesion. Our hypothesis is that global cohesion can better capture *thesis state-*

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2255–2261,
Lisbon, Portugal, 17-21 September 2015.

Corpus	#Essays	Avg.#paras	Avg.#sents	Element distributions							Kappa
				I	P	T	M	S	C	O	
C1	367	7.4	22.7	0.077	0.080	0.087	0.135	0.514	0.095	0.008	0.94
C2	346	7.8	22.5	0.070	0.027	0.069	0.181	0.530	0.114	0.006	0.93
C3	197	9.1	27.7	0.082	–	0.045	0.187	0.571	0.106	0.007	0.91
Avg.	303	7.9	23.7	0.077	0.053	0.067	0.169	0.538	0.105	0.007	0.93

Table 2: Basic statistics of the annotated corpora. Para and sent are short for paragraph and sentence.

ments and help distinguishing them from overwhelming *supporting idea* sentences. Experiments were conducted on essays written by Chinese high school students in the mother tongue. The results confirm our hypothesis. Our method achieves significant improvements of +%5 $F1$ score on *thesis* and *main idea* sentences by adding cohesion features. The features related to global cohesion are most discriminative.

2 Data Annotation

We mainly use the discourse elements defined by Burstein et al. (2003b) except for adding a *prompt* element. The discourse element definitions are listed in Table 1. We asked two labelers from the college of liberal art of a university to conduct data annotation. Provided a detail manual with element definitions, explanations and examples, the labelers assigned a discourse element to each sentence.

We collected three corpora, two of which (C1 and C2) are prompt-directed and one (C3) is prompt-free. All essays were written by Chinese high school students in Chinese. The prompt-directed essays are samples of essays written by senior high school students when they were taking a mock examination. The students were required to write a pervasive essay related to a given prompt. The prompts of corpora C1 and C2 are different. The prompt-free essays in C3 were crawled from an online writing assistance website, where the essays were used as writing examples of persuasive essays written by high school students. The average essay lengths on three corpora are 795, 772 and 864 Chinese characters respectively. The other basic statistics of the annotated corpora are listed in Table 2.

During annotation, the labelers found cases of difficulties about ambiguous elements. For example, content about the prompt and the main thesis can be mentioned in the same sentence. In such cases, *thesis statements* have priority over other elements to be labeled, since identifying *thesis statements* is more important for some potential applications (Burstein et al., 2001).

From each corpus, 100 essays were labeled by both annotators for computing agreements, and the others were labeled independently. The label agreements measured with Kappa (Cohen et al., 1960) are high as shown in Table 2. The disagreements were resolved by discussion. The distributions of discourse elements are also shown in Table 2. We can see that they are imbalanced. The *supporting idea* sentences account for more than 53%, while the *thesis statements* account for only 34% in total. As a result, the distinction between minority *thesis statements* from overwhelming *supporting idea* sentences is a major challenge.

3 Discourse Element Identification

Identifying discourse elements in student essays can be seen as a functional segmentation of discourse (Webber et al., 2011). In this work, we focus on utilizing supervised feature-based machine learning models for this task.

3.1 Learning Models

Discourse element identification can be casted as a classification problem that sentences are classified independently using a classifier, e.g. naive Bayes (Burstein et al., 2001), decision tree (Burstein et al., 2003b) and Support Vector Machines (SVMs) (Stab and Gurevych, 2014). It can also be solved in a sequence labeling framework, which models the whole sentence sequence and captures the correlations among predictions. For example, Conditional Random Fields (CRFs) have be studied for similar task on argumentative zoning of scientific documents (Guo et al., 2011).

We will evaluate different types of features using two representative models respectively: the SVM model and the linear-chain CRF model.

3.2 Basic Features

Before feature extraction, sentence splitting, word segmentation,POS and NE tagging are done using a Chinese language processing toolkit (Che et al., 2010). Most basic features are adapted from previous work (Burstein et al., 2003a; Stab and Gurevych, 2014; Persing et al., 2010). For each sentence, the following feature sets are extracted.

Position features The relative position of its paragraph (first, last or body) in the essay and its relative position (first, last or body) in the paragraph are modeled as a set of binary features. The index of the sentence is also used as a feature.

Indicator features Cue words/phrases like "我认为(in my opinion)" and "总之(in conclusion)" are used as indicators. Partial indicators are adapted from the ones used by Persing et al. (2010). More Chinese specific indicators are then augmented manually. We use a binary feature denoting a reference to the first person ("我(I)","我们(We)") in the sentence. We also use a binary feature to indicate whether the sentence contains a modal verb like "应该(should)" and "希望(hope)".

Lexical features Binary features are modeled for all connectives and adverbs, which are identified based on POS tags.

Structural features The number of words, the number of clauses in the sentence and the number of sentences in the same paragraph are used as features. We also define binary features based on punctuation which indicate whether the sentence ends with a full stop, question mark, exclamation mark or no sentence-final punctuation.

Topic and prompt features For each sentence, the cosine similarities to the essay title and to the prompt are used as features.

4 Identification based on Cohesion

4.1 Cohesive Chains

We mainly exploit reference and lexical cohesion.

Creating identity chains Reference refers to resources for referring to a participant whose identity is recoverable (Schiffrin et al., 2008). We focus on person identities, because person names might be mentioned when describing facts. Firstly, we extract all nouns/entities with a POS/NE tag *person* as identities. Secondly, we conduct a simple third-person pronoun resolution by selecting the nearest proper antecedent identity within the same paragraph. Finally, an identity and all its anaphora together form an identity chain.

Creating lexical chains Lexical cohesion is referred to relations between text using lexical repetition, synonymy or near synonymy. We don't distinguish between systematic semantic relations and non-systematic semantic relations (Berzlánovich et al., 2008) nor use a thesaurus (Hirst and St-Onge, 1998). Instead, we compute the relatedness of two words based on their distributed representations, which are learned using the Word2Vec toolkit (Mikolov et al., 2013). The data for learning word representations consists of student essays and textbooks crawled from the Web. The vocabulary size is about $490k$.

We extract nouns, adjectives and verbs (excluding auxiliary verbs) instead of using nouns only (Hirst and St-Onge, 1998) for constructing lexical chains . We firstly cluster words into clusters in a graph based manner. Each word corresponds to a node in a graph. If the relatedness of two words is larger than a threshold T, they are considered as related and linked by an edge. After constructing all edges in this way, every connected subgraph forms a word cluster. Through the essay, all occurrences of the words from the same cluster form a lexical chain.

We discard identity and lexical chains that exist within single sentences, since they can't capture cohesive relations among sentences.

4.2 Global and Local Sentence Chains

We organize sentences based on cohesive chains. Sentences that contain members from the same identity chain or lexical chain form a sentence chain. The sentence chains represent cohesive relations among sentences.

In persuasive writing, discourse elements are commonly linked globally. For example, *main ideas* are usually related to each other because they are about different aspects of the main thesis, and *thesis* and *conclusion* should echo each other as well. Therefore, we attempt to explicitly categorize sentence chains into local chains and global chains based on subtopic coverage. A local chain represents sentences that share cohesive relations and gather locally within single subtopics. In contrast, a global chain represents sentences with cohesive relations distribute across multiple subtopics. Heuristically, we expect that *thesis statements* can be better captured by global chains, while sentences that state facts or provide evidences are associated to local chains.

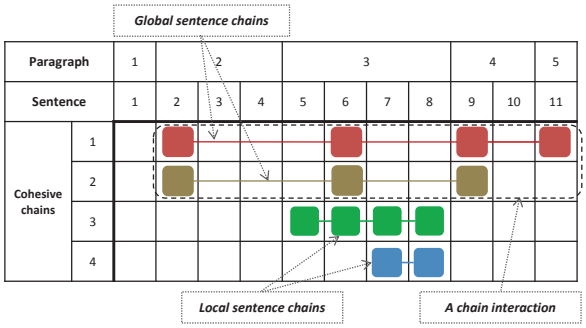

Figure 1: An illustration of global/local sentence chains. Each solid node in the grid indicates that a sentence contains a word from a cohesive chain.

Although subtopics can be identified by existing text segmentation algorithms (Hearst, 1997; Filippova and Strube, 2006), we observe that in student essays a subtopic boundary usually coincides with a paragraph boundary and almost all subtopic segments are within one paragraph and only a few of subtopics are within two or more paragraphs. Therefore, we simply assume that each paragraph corresponds to a subtopic. Based on the assumption, chain classification is approximated based on chain span over paragraphs. A sentence chain is classified as a global chain, if its members appear in at least N paragraphs, otherwise it is classified as a local chain. We set $N = 3$, which means a global chain would cover at least two subtopics considering most subtopics finish within two paragraphs. One sentence can be involved in multiple sentence chains. An illustration of global and local sentence chains is shown in Figure 1.

4.3 Cohesion Features from Sentence Chains

We develop cohesion features for a sentence from the sentence chains that involve it. Such features are beyond the intrinsic properties of the sentence itself but describe relations to other sentences.

Chain-type features We consider four combined types of sentence chains: *global-identity, local-identity, global-lexical* and *local-lexical* chains. The number of each type of chains that involve the sentence is used as a feature.

Global-title feature If the sentence is in a global sentence chain and the corresponding cohesive chain contains a word in the title, a binary feature *global-title* is set as *true*, otherwise set as *false*. Containing globally distributed title words is thought of as an indicator of *thesis statements*.

Interaction features Hasan (1984, 1985) defined that an interaction between two chains takes place when multiple members of a chain relate in the same way to more than one members of another chain, which can be used to distinguish central tokens from peripheral tokens. Hoey (1991) explored similar interactions to assess the centrality of sentences. This indicates that chain interactions might be signals of important content.

Similarly, we say two sentence chains interact with each other, if they have more than one sentence in common. An example is shown in Figure 1. Moreover, if two chains are both global chain, we term it a global interaction, otherwise a local interaction. The shared sentences by two chains are named as *global or local interaction sentences* accordingly. Two binary features are derived: the sentence is or not a *global-interaction* sentence and it is or not a *local-interaction* sentence.

Strength features We attempt to measure the overall strength of the sentence chains that involve the sentence. The features include the number of chains, the maximum and average number of covered sentences and paragraphs over chains, among which the ones related to paragraphs can be seen as measuring the global cohesion strength.

5 Evaluation

5.1 Settings

We evaluated the effectiveness of **Cohesion** features by comparing with the baseline that uses the **Basic** features introduced in Section 3.2.

We adopted $precision(P)$, $recall(R)$ and $F1$-measure($F1$) as evaluation metrics. The threshold T used to determine whether two words are related was set to 0.8 empirically. Because sentences with the discourse element *Other* are few, we didn't evaluate the performance on it.

We conducted experiments on three corpora respectively using 5-fold cross-validation. We compared various SVM classifiers with different kernels implemented in the LibSVM toolkit (Chang and Lin, 2011) and the linear-chain CRF model (Lafferty et al., 2001). When using CRF, the prediction of previous sentence is considered for current sentence. In our experiments, CRF achieves significant superior performance than SVM both when using basic features alone and after adding cohesion features. This indicates that incorporating correlations among sequential predictions are important for this task. Next, we only report the experimental results of using the CRF Model.

Element	Features	C1			C2			C3			
		P	R	$F1$	P	R	$F1$	P	R	$F1$	$avg. \triangle (F1)$
Introduction	Basic	84.5	89.6	86.8	82.2	80.7	81.5	80.6	90.1	85.0	+3.7
	+ Cohesion	87.2	90.8	88.8	85.6	84.8	85.2	87.3	94.4	90.6	
Prompt	Basic	89.7	86.9	88.2	77.2	69.0	72.5	–	–	–	+1.9
	+ Cohesion	91.1	89.2	90.1	82.0	69.1	74.4	–	–	–	
Thesis	Basic	76.5	69.0	72.4	69.9	61.1	64.9	73.3	57.5	64.0	+5.1
	+ Cohesion	78.3	73.1	75.5	75.4	63.8	68.6	77.3	68.9	72.7	
Main idea	Basic	71.4	59.1	64.5	69.0	60.9	64.6	69.4	54.0	60.7	+5.4
	+ Cohesion	75.7	65.3	70.0	73.6	61.3	66.8	75.7	64.3	69.4	
Supporting idea	Basic	86.1	91.4	88.6	83.8	89.6	86.6	83.8	90.5	87.0	+1.8
	+ Cohesion	88.0	92.3	90.1	84.2	91.6	87.7	87.7	92.2	89.9	
Conclusion	Basic	87.2	89.9	88.4	85.6	88.5	87.0	88.1	91.0	89.5	+2.2
	+ Cohesion	89.1	91.9	90.4	86.0	90.7	88.2	92.1	94.0	93.1	

Table 3: Experimental results on six discourse elements over three corpora using the CRF model.

5.2 Experimental Results

The experimental results on three corpora are shown in Table 3. We tested statistical significance for $F1$ scores and found that all improvements were significant with $p < 0.01$ based on the pairwise t-test. We can see that adding cohesion features obtain improvements on all discourse elements over three corpora. Especially, the cohesion features contribute to large improvements of +5.1% and +5.4% average $F1$ score on identifying *thesis* and *main idea* sentences. By analyzing the confusion matrix, we found that the improvements mainly come from more accurately distinguishing *thesis* and *main idea* sentences from *introduction* and *supporting idea* sentences.

We are interested in that which between the local and global cohesion contributes more to distinguish *thesis statements*. To this end, we used Area under the ROC Curve (AUC)(Swets, 1988) to measure the discriminative power of individual features. Larger AUC of a feature indicates better discriminative performance. The experiment was done on the dataset mixing all sentences from three corpora. We divided sentences into *thesis statements* and *non-thesis statements* according to their true element labels. As the results in Table 4 show, global cohesion related features are of higher rank with regard to the discriminative power. Local cohesion relates more to *non-thesis statements*, though it is not so discriminative as global cohesion. The results indicate that separating global cohesion from local cohesion help the distinction between *thesis statements* and others. Features related to identity chains alone don't show much discriminative ability. But they increase the macro $F1$ score by 0.9% in combination with other features.

Cohesion Feature	AUC
Global-lexical	0.712
Avg.#paras	0.670
Global-title	0.664
Max.#para	0.659
Global-interaction	0.654
Max.#sents	0.636
Avg.#sents	0.613
#Chains	0.601
Local-title	0.522
Global-identity	0.510
Local-identity	0.481
Local-interaction	0.476
Local-lexical	0.431

Table 4: Discriminative powers of individual features by Areas under ROC curve (AUC).

6 Conclusion

We have investigated the impact of cohesion for identifying discourse elements in student essays. Our method creates sentence chains by means of cohesive resources and separates global chains from local ones based on the subtopic coverage. New features for each sentence are derived from the properties of the sentence chains involving it. Experimental results show the effectiveness of cohesion features and the discriminative ability of global cohesion for identifying *thesis statements*.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No.61402304), the Beijing Municipal Natural Science Foundation (No.4154065) and the Humanity & Social Science General Project of Ministry of Education (No.14YJAZH046).

References

Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. *Advances in automatic text summarization*, pages 111–121, 1999.

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

Ildikó Berzlánovich, Markus Egg, and Gisela Redeker. Coherence structure and lexical cohesion in expository and persuasive texts. In *Proceedings of the Workshop Constraints in Discourse III*, pages 19–26, 2008.

Jill Burstein, Daniel Marcu, Slava Andreyev, and Martin Chodorow. Towards automatic classification of discourse elements in essays. In *Proceedings of the 39th annual Meeting on Association for Computational Linguistics*, pages 98–105. Association for Computational Linguistics, 2001.

Jill Burstein, Martin Chodorow, and Claudia Leacock. Criterionsm online essay evaluation: An application for automated evaluation of student essays. In *IAAI*, pages 3–10, 2003a.

Jill Burstein, Daniel Marcu, and Kevin Knight. Finding the write stuff: Automatic identification of discourse structure in student essays. *Intelligent Systems, IEEE*, 18(1):32–39, 2003b.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

Wanxiang Che, Zhenghua Li, and Ting Liu. Ltp: A chinese language technology platform. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 13–16. Association for Computational Linguistics, 2010.

Jacob Cohen et al. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Katja Filippova and Michael Strube. Using linguistically motivated features for paragraph boundary identification. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 267–274. Association for Computational Linguistics, 2006.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 562–569. Association for Computational Linguistics, 2003.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 273–283. Association for Computational Linguistics, 2011.

MAK Halliday and Ruqaiya Hasan. Cohesion in english (english language). *London, 1976; Martin JR*, 1976.

Ruqaiya Hasan. Coherence and cohesive harmony. *Understanding Reading Comprehesion: Cognition, Lanugage and The Structure of Prose*, 1984.

Ruqaiya Hasan. The texture of a text. *Language, Context and Text*, 1985.

Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, pages 33–64, 1997.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. Evaluating multiple aspects of coherence in student essays. In *HLT-NAACL*, pages 185–192, 2004.

Graeme Hirst and David St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 305:305–332, 1998.

Michael Hoey. Patterns of lexis in text. 1991.

Pei-Yun Hsueh, Johanna D Moore, and Steve Renals. Automatic segmentation of multiparty dialogue. In *EACL*, 2006.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289. Morgan Kaufmann, 2001.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48, 1991.

Isaac Persing, Alan Davis, and Vincent Ng. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239. Association for Computational Linguistics, 2010.

Deborah Schiffrin, Deborah Tannen, and Heidi E Hamilton. *The handbook of discourse analysis*. John Wiley & Sons, 2008.

Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In *EMNLP 2014*, pages 46–56, 2014.

John A Swets. Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, 1988.

Bonnie Webber, Markus Egg, and Valia Kordoni. Discourse structure and language technology. *Natural Language Engineering*, 18(4): 437–490, 2011.