Discourse Mode Identification in Essays

Wei Song[†], **Dong Wang**[‡], **Ruiji Fu**[‡], **Lizhen Liu**[†], **Ting Liu**[§], **Guoping Hu**[‡]
[†]Information Engineering, Capital Normal University, Beijing, China
[‡]iFLYTEK Research, Beijing, China
[§]Harbin Institute of Technology, Harbin, China
{wsong, lzliu}@cnu.edu.cn, {dongwang4,rjfu, gphu}@iflytek.com, tliu@ir.hit.edu.cn

Abstract

Discourse modes play an important role in writing composition and evaluation. This paper presents a study on the manual and automatic identification of *narration*, *exposition*, *description*, *argument* and *emotion expressing* sentences in narrative essays. We annotate a corpus to study the characteristics of discourse modes and describe a neural sequence labeling model for identification. Evaluation results show that discourse modes can be identified automatically with an average F1-score of 0.7. We further demonstrate that discourse modes can be used as features that improve automatic essay scoring (AES). The impacts of discourse modes for AES are also discussed.

1 Introduction

Discourse modes, also known as rhetorical modes, describe the purpose and conventions of the main kinds of language based communication.Most common discourse modes include narration, description, exposition and argument. A typical text would make use of all the modes, although in a given one there will often be a main mode. Despite their importance in writing composition and assessment (Braddock et al., 1963), there is relatively little work on analyzing discourse modes based on computational models. We aim to contribute for automatic discourse mode identification and its application on writing assessment.

The use of discourse modes is important in writing composition, because they relate to several aspects that would influence the quality of a text.

First, discourse modes reflect the organization of a text. Natural language texts consist of sentences which form a unified whole and make up the discourse (Clark et al., 2013). Recognizing the structure of text organization is a key part for discourse analysis. Meurer (2002) points that discourse modes stand for unity as they constitute general patterns of language organization strategically used by the writer. Smith (2003) also proposes to study discourse passages from a linguistic view of point through discourse modes. The organization of a text can be realized by segmenting text into passages according to the set of discourse modes that are used to indicate the functional relationship between the several parts of the text. For example, the writer can present major events through narration, provide details with description and establish ideas with argument. The combination and interaction of various discourse modes make an organized unified text.

Second, discourse modes have rhetorical significance. Discourse modes are closely related to rhetoric (Connors, 1981; Brooks and Warren, 1958), which offers a principle for learning how to express material in the best way. Discourse modes have different preferences on expressive styles. Narration mainly controls story progression by introducing and connecting events; exposition is to instruct or explain so that the language should be precise and informative; argument is used to convince or persuade through logical and inspiring statements; description attempts to bring detailed observations of people and scenery, which is related to the writing of figurative language; the way to express emotions may relate to the use of rhetorical devices and poetic language. Discourse modes reflect the variety of expressive styles. The flexible use of various discourse modes should be important evidence of language proficiency.

According to the above thought, we propose the discourse mode identification task. In particular, we make the following contributions:

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 112–122
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-1011

- We build a corpus of narrative essays written by Chinese students in native language. Sentence level discourse modes are annotated with acceptable inter-annotator agreement. Corpus analysis reveals the characteristics of discourse modes in several aspects, including discourse mode distribution, co-occurrence and transition patterns.

- We describe a multi-label neural sequence labeling approach for discourse mode identification so that the co-occurrence and transition preferences can be captured. Experimental results show that discourse modes can be identified with an average F1-score of 0.7, indicating that automatic discourse mode identification is feasible.

- We demonstrate the effectiveness of taking discourse modes into account for automatic essay scoring. A higher ratio of description and emotion expressing can indicate essay quality to a certain extent. Discourse modes can be potentially used as features for other NLP applications.

2 Related Work

2.1 Discourse Analysis

Discourse analysis is an important subfield of natural language processing (Webber et al., 2011). Discourse is expected to be both cohesive and coherent. Many principles are proposed for discourse analysis, such as coherence relations (Hobbs, 1979; Mann and Thompson, 1988), the centering theory for local coherence (Grosz et al., 1995) and topic-based text segmentation (Hearst, 1997). In some domains, discourse can be segmented according to specific discourse elements (Hutchins, 1977; Teufel and Moens, 2002; Burstein et al., 2003; Clerehan and Buchbinder, 2006; Song et al., 2015).

This paper focuses on discourse modes influenced by Smith (2003). From the linguistic view of point, discourse modes are supposed to have different distributions of situation entity types such as event, state and generic (Smith, 2003; Mavridou et al., 2015). Therefore, there is work on automatically labeling clause level situation entity types (Palmer et al., 2007; Friedrich et al., 2016). Actually, situation entity type identification is also a challenging problem. It is even harder for processing Chinese language, since Chinese doesn't have grammatical tense (Xue and Zhang, 2014) and sentence components are often omitted. This increases the difficulties for situation entity type based discourse mode identification. In this paper, we investigate an end-to-end approach to directly model discourse modes without the necessity of identifying situation entity types first.

2.2 Automatic Writing Assessment

Automatic writing assessment is an important application of natural language processing. The task aims to let computers have the ability to appreciate and criticize writing. It would be hugely beneficial for applications like automatic essay scoring (AES) and content recommendation.

AES is the task of building a computer-aided scoring system, in order to reduce the involvement of human raters. Traditional approaches are based on supervised learning with designed feature templates (Larkey, 1998; Burstein, 2003; Attali and Burstein, 2006; Chen and He, 2013; Phandi et al., 2015; Cummins et al., 2016). Recently, automatic feature learning based on neural networks starts to draw attentions (Alikaniotis et al., 2016; Dong and Zhang, 2016; Taghipour and Ng, 2016).

Writing assessment involves highly technical aspects of language and discourse. In addition to give a score, it would be better to provide explainable feedbacks to learners at the same time. Some work has studied several aspects such as spelling errors (Brill and Moore, 2000), grammar errors (Rozovskaya and Roth, 2010), coherence (Barzilay and Lapata, 2008), organization of argumentative essays (Persing et al., 2010) and the use of figurative language (Louis and Nenkova, 2013). This paper extends this line of work by taking discourse modes into account.

2.3 Neural Sequence Modeling

A main challenge of discourse analysis is hard to collect large scale data due to its complexity, which may lead to data sparseness problem. Recently, neural networks become popular for natural language processing (Bengio et al., 2003; Collobert et al., 2011). One of the advantages is the ability of automatic representation learning. Representing words or relations with continuous vectors (Mikolov et al., 2013; Ji and Eisenstein, 2014) embeds semantics in the same space, which benefits alleviating the data sparseness problem and enables end-to-end and multi-task learning. Recurrent neural networks (RNNs) (Graves, 2012) and the variants like Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent (GRU) (Cho et al., 2014) neural networks show good performance for capturing long distance dependencies on tasks like Named Entity Recognition (NER) (Chiu and Nichols, 2016; Ma and Hovy, 2016), dependency parsing (Dyer et al., 2015) and semantic composition of documents (Tang et al., 2015). This work describes a hierarchical neural architecture with multiple label outputs for modeling the discourse mode sequence of sentences.

3 Discourse Mode Annotation

We are interested in the use of discourse modes in writing composition. This section describes the discourse modes we are going to study, an annotated corpus of student essays and what we learn from corpus analysis.

3.1 Discourse Modes

Discourse modes have several taxonomies in the literature. Four basic discourse modes are *narration*, *description*, *exposition* and *argument* in English composition and rhetoric (Bain, 1890). Smith (2003) proposes five modes for studying discourse passages: *narrative*, *description*, *report*, *information* and *argument*. In Chinese composition, discourse modes are categorized into *narration*, *description*, *exposition*, *argument* and *emotion expressing* (Zhu, 1983).

These taxonomies are similar. Their elements can mostly find corresponding ones in other taxonomies literally or conceptually, e.g., exposition mode has similar functions to information mode. Emotion expressing that is to express the writer's emotions is relatively special. It can be realized by expressing directly or through lyrical writing with beautiful and poetic language. It is also related to appeal to emotion, which is a method for argumentation by the manipulation of the recipient's emotions in classical rhetoric (Aristotle and Kennedy, 2006). Proper emotion expressing can touch the hearts of the readers and improve the expressiveness of writing. Therefore, considering it as an independent mode is also reasonable.

We cope with essays written in Chinese in this work so that we follow the Chinese convention with five discourse modes. Emotion expressing is added on the basis of four recognized discourse modes and Smith's report mode is viewed as a subtype of description mode: *dialogue description*.

In summary, we study the following discourse modes:

- **Narration** introduces an event or series of events into the universe of discourse. The events are temporally related according to narrative time.
  E.g., *Last year, we drove to San Francisco along the State Route 1 (SR 1).*

- **Exposition** has a function to explain or instruct. It provides background information in narrative context. The information presented should be general and (expected to be) well accepted truth.
  E.g., *SR 1 is a major north-south state highway that runs along most of the Pacific coastline of the U.S.*

- **Description** re-creates, invents, or vividly show what things are like according to the five senses so that the reader can picture that which is being described.
  E.g., *Along SR 1 are stunning rugged coastline, coastal forests and cliffs, beautiful little towns and some of the West coast's most amazing nature.*

- **Argument** makes a point of view and proves its validity towards a topic in order to convince or persuade the reader.
  E.g., *Point Arena Lighthouse is a must see along SR 1, in my opinion.*

- **Emotion expressing**[1] presents the writer's emotions, usually in a subjective, personal and lyrical way, to involve the reader to experience the same situations and to be touched.
  E.g., *I really love the ocean, the coastline and all the amazing scenery along the route. When could I come back again?*

The distinction between discourse modes is expected to be clarified conceptually by considering their different communication purposes. However, there would still be specific ambiguous and vague cases. We will describe the data annotation and corpus analysis in the following parts.

[1] In some cases, we use emotion for short.

	INITIAL			FINAL		
	P	R	F	P	R	F
Nar	0.90	0.88	0.89	0.96	0.84	0.90
Exp	0.79	0.73	0.76	0.89	0.76	0.81
Des	0.84	0.74	0.79	0.87	0.65	0.74
Emo	0.75	0.68	0.71	0.79	0.73	0.76
Arg	0.35	0.28	0.31	0.76	0.61	0.68
Avg.	0.73	0.66	0.69	0.87	0.71	0.78
κ		0.55			0.72	

Table 1: Inter-annotator agreement between two annotators on the dominant discourse mode. Initial: The result of the first round annotation; Final: The result of the final annotation; κ: Agreement measured with Cohen's Kappa.

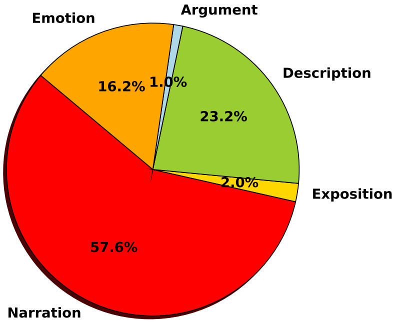


Figure 1: The distribution of dominant modes.

3.2 Data Annotation

Discourse modes are almost never found in a pure form but are embedded one within another to help the writer achieve the purpose, but the emphasis varies in different types of writing. We focus on narrative essays. A good narrative composition must properly manipulate multiple discourse modes to make it vivid and impressive.

The corpus has 415 narrative essays written by high school students in their native Chinese language.The average number of sentences is 32 and the average length is 670 words.

We invited two high school teachers to annotate discourse modes at sentence level, expecting their background help for annotation. A detail manual was discussed before annotation.

We notice that discourse modes can mix in the same sentence. Therefore, the annotation standard allows that one sentence can have multiple modes. But we require that every sentence should have a *dominant mode*. The annotators should try to think in the writer's perspective and guess the writer's main purpose of writing the sentence in order to decide the dominant mode.

Among the discourse modes, description can be applied in various situations. We focus on the following description types: portrait, appearance, action, dialogue, psychological, environment and detail description. If a sentence has any type of description, it would be assigned a description label.

3.3 Corpus Analysis

We conducted corpus analysis on the annotated data to gain observations on several aspects.

Inter-Annotator Agreement: 50 essays were independently annotated by two annotators. We evaluate the inter-annotator agreement on the dominant mode. The two annotators' annotations are used as the golden answer and prediction respectively. We compute the precision, recall and F1-score for each discourse mode separately to measure the inter-annotator agreement. Precision and recall are symmetric for the two annotators.

The result of the first round annotation is shown in the INITIAL columns of Table 1. The agreement on argument mode is low, while the agreement on other modes is acceptable. The average F1-score is 0.69. The Cohen's Kappa (Cohen et al., 1960) is 0.55 over all judgements on the dominant mode.

The main disagreement on argument lies in the confusion with emotion expressing. Consider the following sentence:

> *Father's love is the fire that lights the lamp of hope.*

One annotator thought that it is expressed in an emotional and lyrical way so that the discourse mode should be emotion expressing. The other one thought that it (implicitly) gives a point and should be an argument. Many disagreements happened in cases like this.

Based on the observations of the first round annotation, we discussed and updated the manual and let the annotators rechecked their annotations. The final result is shown in the FINAL columns of Table 1. The agreement on description decreases. Annotators seem to be more conservative on labeling description as the dominant mode. The overall average F1-score increases to 0.78 and the Cohen's Kappa is 0.72. This indicates that humans can reach an acceptable agreement on the dominant discourse mode of sentences after training.

Discourse mode distribution: After the training phase, the annotators labeled the whole corpus. Figure 1 shows the distribution of dominant

Mode	Nar	Exp	Des	Emo	Arg
Nar	5285	11	2552	65	2
Exp	-	148	11	1	1
Des	-	-	2538	105	8
Emo	-	-	-	1947	63
Arg	-	-	-	-	318

Table 2: Co-occurrence of discourse modes in the same sentences. The numbers in diagonal indicate the number of sentences with a single mode.

from \ to	Nar	Exp	Des	Emo	Arg
Nar	72%	-	17%	7%	1%
Exp	59%	8%	8%	16%	6%
Des	42%	-	53%	3%	-
Emo	25%	2%	4%	66%	1%
Arg	27%	-	4%	12%	54%
Begin with	50%	3%	6%	32%	7%
End with	12%	1%	2%	76%	6%

Table 3: Transition between discourse modes of consecutive sentences and the distribution of discourse modes that essays begin with and end with.

discourse modes. The distribution is imbalanced. Narration, description and emotion expressing are the main discourse modes in narrative essays, while exposition and argument are rare.

Co-occurrence: Statistics show that 78% of sentences have only one discourse mode, and 19% have two discourse modes, and 3% have more than two discourse modes.

Table 2 shows the co-occurrence of discourse modes. The numbers that are in the diagonal represent the distribution of discourse modes of sentences with only one mode. The numbers that are not in the diagonal indicate the co-occurrence of modes in the same sentences. We can see that description tends to co-occur with narration and emotion expressing. Description can provide states that happen together with events and emotion-evoking scenes are often described to elicit a strong emotional response, for example:

> *The bright moon hanging on the distant sky reminds me of my hometown miles away.*

Emotion expressing and argument also co-occur in some cases. It is reasonable, since a successful emotional appeal can enhance the effectiveness of an argument.

Generally, these observations are consistent with intuition. Properly combining multiple modes could produce impressive sentences.

Transition: Table 3 shows the transition matrix between the dominant modes of consecutive sentences within the same paragraphs. All modes tend to transit to themselves except exposition, which is rare and usually brief. This means that discourse modes of adjacent sentences have high correlation. We also see that narration and emotion are more often at the beginning and the end of essays. The above observations indicate that discourse modes have local preferred patterns.

To summarize, the implications of corpus analysis include: (1) Manual identification of discourse modes is feasible with an acceptable inter-annotator agreement; (2) The distribution of discourse modes in narrative essays is imbalanced; (3) About 22% sentences have multiple discourse modes; (4) Discourse modes have local transition patterns that consecutive discourse modes have high correlation.

4 Discourse Mode Identification based on Neural Sequence Labeling

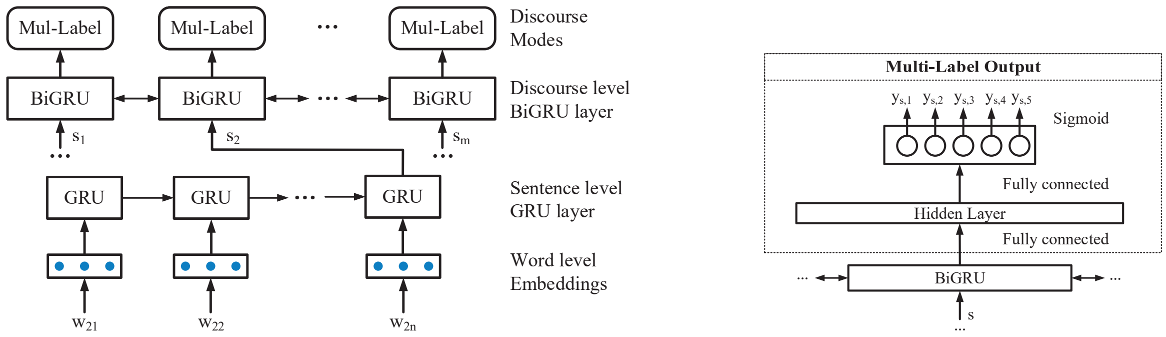
This section describes the proposed method for discourse mode identification. According to the corpus analysis, sentences often have multiple discourse modes and prefer local transition patterns. Therefore, we view this task as a multi-label sequence labeling problem.

4.1 Model

We propose a hierarchical neural sequence labeling model to capture multiple level information. Figure 2(a) shows the basic architecture. We introduce it from the bottom up.

Word level embedding layer: We transform words into continuous vectors, word embeddings. Vector representation of words is useful for capturing semantic relatedness. This should be effective in our case, since large amount of training data is not available. It is unrealistic to learn the embedding parameters on limited data so that we just look up embeddings of words from a pre-trained word embedding table. The pre-trained word embeddings were learned with the Word2Vec toolkit (Mikolov et al., 2013) on a domain corpus which consists of about 490,000 student essays. The embeddings are kept unchanged during learning and prediction.

Sentence level GRU layer: Each sentence is a sequence of words. We feed the word embeddings into a forward recurrent neural networks. Here,



(a) The basic hierarchical architecture.

(b) The detail of the Mul-Label layer

Figure 2: The multi-label neural sequence labeling model for discourse mode identification.

we use the GRU (Cho et al., 2014) as the recurrent unit. The GRU is to make each recurrent unit to adaptively capture dependencies of different time scales. The output of the last time-step is used as the representation of a sentence.

Discourse level bidirectional-GRU layer: An essay consists of a sequence of sentences. Accessing information of past and future sentences provides more contextual information for current prediction. Therefore, we use a bidirectional RNN to connect sentences. We use the GRU as the recurrent unit, which is also shown effective on semantic composition of documents for sentiment classification (Tang et al., 2015). The BiGRU represents the concatenation of the hidden states of the forward GRU and the backward GRU units.

Multi-Label layer: Since one sentence can have more than one discourse mode, our model allows multiple label outputs. Figure 2(b) details the Mul-Label layer in Figure 2(a). The representation of each sentence after the bidirectional-GRU layer is first fully connected to a hidden layer. The hidden layer output is then fully connected to a five-way output layer, corresponding to five discourse modes. The sigmoid activation function is applied to each way to get the probability that whether corresponding discourse mode should be assigned to the sentence.

In the training phase, the probability of any labeled discourse modes is set to 1 and the others are set to 0. In the prediction phase, if the predicted probability of a discourse mode is larger than 0.5, the discourse mode would be assigned.

4.1.1 Considering Paragraph Boundaries

Different from NER that processes a single sentence each time, our task processes sequences of sentences in discourse, which are usually grouped by paragraphs to split the whole discourse into several relatively independent segments. Sentences from different paragraphs should have less effect to each other, even though they are adjacent.

To capture paragraph boundary information, we insert an empty sentence at the end of every paragraph to indicate a paragraph boundary. The empty sentence is represented by a zero vector and its outputs are set to zeros as well. We expect this modification can better capture position related information.

4.2 Implementation Details

We implement the model using the Keras library.[2] The models are trained with the binary cross-entropy objective. The optimizer is Adam (Kingma and Ba, 2014). The word embedding dimension is 50. The dimension of the hidden layer in Mul-Label layer is 100. The length of sentences is fixed as 40. All other parameters are set by default parameter values. We adopt early stopping strategy (Caruana et al., 2000) to decide when the training process stops.

4.3 Evaluation

4.3.1 Data

We use 100 essays as the test data. The remaining ones are used as the training data. 10% of the shuffled training data is used for validation.

4.3.2 Comparisons

We compare the following systems:

- SVM: We use bag of ngram (unigram and bigram) features to train a support vector classifier for sentence classification.

[2] https://github.com/fchollet/keras/

- CNN: We implement a convolutional neural network (CNN) based method (Kim, 2014), as it is the state-of-the-art for sentence classification.

- GRU: We use the sentence level representation in Figure 2(a) for sentence classification.

- GRU-GRU(GG): This method is introduced in this paper in §4.1, but it doesn't consider paragraph information.

- GRU-GRU-SEG (GG-SEG): The model considers paragraph information on the top of G-G as introduced in §4.1.1.

The first three classification based methods classify sentences independently. To deal with multiple labels, the classifiers are trained for each discourse mode separately. At prediction time, if the classifier for any discourse mode predicts a sentence as positive, the corresponding discourse mode would be assigned.

4.3.3 Evaluation Results

Table 4 shows the experimental results. We evaluate the systems for each discourse mode with F1-score, which is the harmonic mean of precision and recall. The best performance is in bold.

The SVM performs worst among all systems. The reason is due to the data sparseness and term-mismatch problem, since the size of the annotated dataset is not big enough. In contrast, systems based on neural networks with pre-trained word embeddings achieve much better performance.

The CNN and GRU have comparable performance. The GRU is slightly better. The two methods don't consider the semantic representations of adjacent sentences.

The GG and GG-SEG explore the semantic information of sentences in a sequence by the bidirectional GRU layer. The results demonstrate that considering such information improve the performance on all discourse modes. This proves the advantage of sequential identification compared with isolated sentence classification.

We can see that the GG-SEG further improves the performance on three minority discourse modes compared with GG. This means that the minority modes may have stronger preference to special locations. Exposition benefits most, since many exposition sentences in our dataset are isolated.

Model \ Mode	Nar	Des	Emo	Arg	Exp
SVM	0.672	0.588	0.407	0.152	0.095
CNN	0.793	0.764	0.594	0.333	0.293
GRU	0.800	0.784	0.615	0.402	0.364
GG	**0.822**	**0.797**	0.680	0.423	0.481
GG-SEG	0.815	0.791	**0.717**	**0.483**	**0.667**

Table 4: The F1-scores of systems on each discourse mode.

The performance on argument is not so good. As we discussed in corpus analysis, argument and emotion expressing mode interact frequently. Because the amount of emotion expressing sentences is much more, distinguishing argument from them is hard. Actually, their functions in narrative essays seem to be similar that both are to deepen the author's response or evoke the reader's response to the story.

The overall average F1-score can reach to 0.7 and the performance on identifying three most common discourse modes are consistent, with an average F1-score above 0.76 using the proposed neural sequence labeling models. Automatic discourse mode identification should be feasible.

5 Essay Scoring with Discourse Modes

Discourse mode identification can potentially provide features for downstream NLP applications. This section describes our attempt to explore discourse modes for automatic essay scoring (AES).

5.1 Essay Scoring Framework

We adopt the standard regression framework for essay scoring. We use support vector regression (SVR) and Bayesian linear ridge regression (BLR-R), which are used in recent work (Phandi et al., 2015). The key is to design effective features.

5.2 Features

The basic feature sets are based on (Phandi et al., 2015).The original feature sets include:

- Length features

- Part-Of-Speech (POS) features

- Prompt features

- Bag of words features

We re-implement the feature extractors exactly according to the description in (Phandi et al., 2015) except for the POS features, since we don't

			Score	
Prompt	#Essays	Avg. len	Range	Median
1	4000	628	0-60	46
2	4000	660	0-50	41
3	3300	642	0-50	41

Table 5: Details of the three datasets for AES.

have correct POS ngrams for Chinese. We complement two additional features: (1) The number of words occur in Chinese Proficiency Test 6 vocabulary; (2) The number of Chinese idioms used.

We further design discourse mode related features for each essay:

- Mode ratio: For each discourse mode, we compute its mode ratio according to $ratio = \frac{\#\text{sentences with the discourse mode}}{\#\text{sentences in the essay}}$. Such features indicate the distribution of discourse modes.

- Bag of ngrams of discourse modes: We use the number of unigrams and bigrams of the dominant discourse modes of the sequence of sentences in the essay as features.

5.3 Experimental Settings

The experiments were conducted on narrative essays written by Chinese middle school students in native language during regional tests. There are three prompts and students are required to write an essay related to the given prompt with no less than 600 Chinese characters. All these essays were evaluated by professional teachers.

We randomly sampled essays from each prompt for experiments. Table 5 shows the details of the datasets. We ran experiments on each prompt dataset respectively by 5-fold cross-validation.

The GG-SEG model was used to identify discourse modes of sentences. Notice that a sentence can have multiple discourse modes. The mode ratio features are computed for each mode separately. When extracting the bag of ngrams of discourse modes features, the discourse mode with highest prediction probability was chosen as the dominant discourse mode.

We use the Quadratic Weighted Kappa (QWK) as the evaluation metric.

5.4 Evaluation Results

Table 6 shows the evaluation results of AES on three datasets. We can see that the BLRR algorithm performs better than the SVR algorithm. No matter which algorithm is adopted, adding discourse mode features make positive contributions for AES compared with using basic feature sets. The trends are consistent over all three datasets.

	QWK Score		
Prompt	1	2	3
SVR-Basic	0.554	0.468	0.457
+ mode	0.6	0.501	0.481
BLRR-Basic	0.683	0.557	0.513
+ mode	**0.696**	**0.565**	**0.527**

Table 6: Evaluation results of AES on three datasets. Basic: the basic feature sets; mode: discourse mode features.

Prompt	1	2	3	Avg
LEN	0.59	0.52	0.45	0.52
Des	0.23	0.24	0.24	0.24
Emo	0.09	0.15	0.12	0.12
Exp	-0.07	0.01	0.01	-0.03
Arg	-0.08	-0.06	-0.1	-0.08
Nar	-0.11	-0.15	-0.12	-0.13

Table 7: Pearson correlation coefficients of mode ratio to essay score. LEN represents essay length.

Impact of discourse mode ratio on scores: We are interested in which discourse mode correlates to essay scores best. Table 7 shows the Pearson correlation coefficient between the mode ratio and essay score. LEN represents the correlation of essay length and is listed as a reference. We can see that the ratio of narration has a negative correlation, which means just narrating stories without auxiliary discourse modes would lead to poor scores. The description mode ratio has the strongest positive correlation to essay scores. This may indicate that using vivid language to provide detail information is essential in writing narrative essays. Emotion expressing also has a positive correlation. It is reasonable since emotional writing can involve readers into the stories. The ratio of argument shows a negative correlation. The reason may be that: first, the identification of argument is not good enough; second, the existence of an argument doesn't mean the quality of argumentation is good. Exposition has little effect on essay scores.

Generally, the distribution of discourse modes shows correlations to the quality of essays. This may relate to the difficulties of manipulating different discourse modes. It is easy for students to use narration, but it is more difficult to manipulate description and emotion expressing well. As a result, the ability of descriptive and emotional writ-

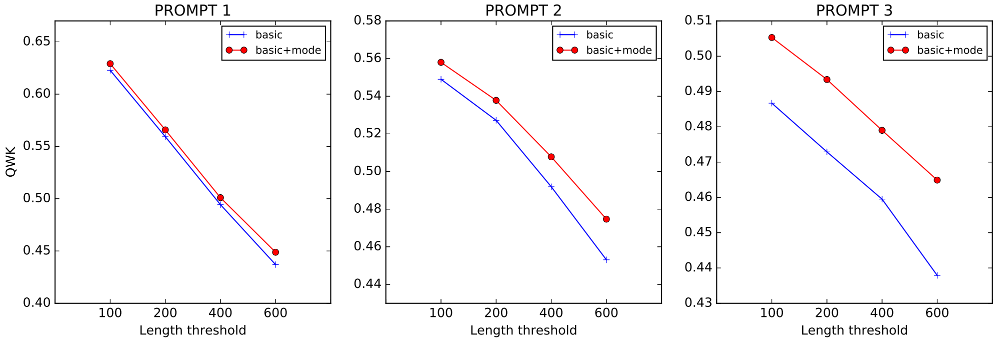


Figure 3: QWK scores on essays satisfying different length thresholds on three prompts. Basic: the basic feature sets; mode: discourse mode features.

ing should be an indicator of language proficiency and can better distinguish the quality of writing.

Impact on scoring essays with various length: It is easy to understand that length is a strong indicator for essay scoring. It is interesting to study that when the effect of length becomes weaker, e.g., the lengths of essays are close, how does the performance of the AES system change?

We conducted experiments on essays with various lengths. Only essays that the length is no less than a given threshold are selected for evaluation. The threshold is set to 100, 200, 400 and 600 Chinese characters respectively. We ran 5-fold cross-validation with BLRR on the datasets after essay selection.

Figure 3 shows the results on three datasets. We can see the following trends: (1) The QWK scores decrease along with shorter essays are removed gradually; (2) Adding discourse mode features always improves the performance; (3) As the threshold becomes larger, the improvements by adding discourse mode features become larger.

The results indicate that the current AES system can achieve a high correlation score when the lengths of essays differ obviously. Even the simple features like length can judge that short essays tend to have low scores. However, when the lengths of essays are close, AES would face greater challenges, because it is required to deeper understand the properties of well written essays. In such situations, features that can model more advanced aspects of writing, such as discourse modes, should play a more important role. It should be also essential for evaluating essays written in the native language of the writer, when spelling and grammar are not big issues any more.

6 Conclusion

This paper has introduced a fundamental but less studied task in NLP—discourse mode identification, which is designed in this work to automatically identify five discourse modes in essays.

A corpus of narrative student essays was manually annotated with discourse modes at sentence level, with acceptable inter-annotator agreement. The corpus analysis revealed several aspects of characteristics of discourse modes including the distribution, co-occurrence and transition patterns.

Considering these characteristics, we proposed a neural sequence labeling approach for identifying discourse modes. The experimental results demonstrate that automatic discourse mode identification is feasible.

We evaluated discourse mode features for automatic essay scoring and draw preliminary observations. Discourse mode features can make positive contributions, especially in challenging situations when simple surface features don't work well. The ratio of description and emotion expressing is shown to be positively correlated to essay scores.

In future, we plan to exploit discourse mode identification for providing novel features for more downstream NLP applications.

Acknowledgements

The research work is partially funded by the National High Technology Research and Development Program (863 Program) of China (No.2015AA015409), National Natural Science Foundation of China (No.61402304), Ministry of Education (No.14YJAZH046), Beijing Municipal Education Commission (KM201610028015, Connotation Development) and Beijing Advanced Innovation Center for Imaging Technology.

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of ACL 2016*. pages 715–725.

Omer Aristotle and George A Kennedy. 2006. *On rhetoric: A theory of civic discourse*. Oxford University Press.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment* 4(3).

Alexander Bain. 1890. *English composition and rhetoric*. Longmans, Green & Company.

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

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3(Feb):1137–1155.

Richard Reed Braddock, Richard Lloyd-Jones, and Lowell Schoer. 1963. *Research in written composition*. JSTOR.

Eric Brill and Robert C Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of ACL 2000*. pages 286–293.

Cleanth Brooks and Robert Penn Warren. 1958. *Modern rhetoric*. Harcourt, Brace.

Jill Burstein. 2003. The e-rater® scoring engine: Automated essay scoring with natural language processing. .

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

Rich Caruana, Steve Lawrence, and Lee Giles. 2000. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Proceedings of NIPS 2000*. pages 402–408.

Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of EMNLP 2013*. pages 1741–1752.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics* 4:357–370.

Kyunghyun Cho, Bart van Merriënboer Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of EMNLP 2014*. pages 1724–1734.

Alexander Clark, Chris Fox, and Shalom Lappin. 2013. *The handbook of computational linguistics and natural language processing*. John Wiley & Sons.

Rosemary Clerehan and Rachelle Buchbinder. 2006. Toward a more valid account of functional text quality: The case of the patient information leaflet. *Text & Talk-An Interdisciplinary Journal of Language, Discourse Communication Studies* 26(1):39–68.

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

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

Robert J Connors. 1981. The rise and fall of the modes of discourse. *College Composition and Communication* 32(4):444–455.

Ronan Cummins, Meng Zhang, and Ted Briscoe. 2016. Constrained multi-task learning for automated essay scoring. In *Proceedings of ACL 2016*. pages 789–799.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In *Proceedings of EMNLP 2016*. pages 1072–1077.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of ACL 2015*. pages 334–343.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of ACL 2016*. pages 1757–1768.

Alex Graves. 2012. Supervised sequence labelling. In *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer Berlin Heidelberg, pages 5–13.

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.

Marti A Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics* 23(1):33–64.

Jerry R Hobbs. 1979. Coherence and coreference. *Cognitive science* 3(1):67–90.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

John Hutchins. 1977. On the structure of scientific texts. *UEA Papers in Linguistics* 5(3):18–39.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of ACL 2014*. pages 13–24.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP 2014*. pages 1746–1751.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Leah S Larkey. 1998. Automatic essay grading using text categorization techniques. In *Proceedings of SIGIR 1998*. pages 90–95.

Annie Louis and Ani Nenkova. 2013. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics* 1:341–352.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of ACL 2016*. pages 1064–1074.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse* 8(3):243–281.

Kleio-Isidora Mavridou, Annemarie Friedrich, Melissa Peate Sørensen, Alexis Palmer, and Manfred Pinkal. 2015. Linking discourse modes and situation entity types in a cross-linguistic corpus study. In *Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*. page 12.

José Luiz Meurer. 2002. Genre as diversity, and rhetorical mode as unity in language use. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies* (43):061–082.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS 2013*. pages 3111–3119.

Alexis Palmer, Elias Ponvert, Jason Baldridge, and Carlota Smith. 2007. A sequencing model for situation entity classification. In *Proceedings of ACL 2007*. pages 896–903.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of EMNLP 2010*. pages 229–239.

Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of EMNLP 2015*. pages 431–439.

Alla Rozovskaya and Dan Roth. 2010. Generating confusion sets for context-sensitive error correction. In *Proceedings of EMNLP 2010*. pages 961–970.

Carlota S Smith. 2003. *Modes of discourse: The local structure of texts*, volume 103. Cambridge University Press.

Wei Song, Ruiji Fu, Lizhen Liu, and Ting Liu. 2015. Discourse element identification in student essays based on global and local cohesion. In *Proceedings of EMNLP 2015*. pages 2255–2261.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of EMNLP 2016*. pages 1882–1891.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of EMNLP 2015*. pages 1422–1432.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics* 28(4):409–445.

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

Nianwen Xue and Yuchen Zhang. 2014. Buy one get one free: Distant annotation of chinese tense, event type and modality. In *Proceedings of LREC 2014*. pages 1412–1416.

Boshi Zhu. 1983. *写作概论(An Introduction to Writing)*. Wuhan, Hubei Educational Press.