Research on Discourse Coherence based on the Analysis Model of Event Chain from the Perspective of Computational Linguistics

Min Fan[1,*], Shanwen Xu[1,**]

[1]College of Foreign Language, Bengbu University, Bengbu, China

Abstract—With the rapid development of network technology, natural language processing has also entered a boom period. Probability and data-driven methods have been widely used in natural language processing. The need for people to extract and retrieve information from the Internet is also increasing, and more and more researchers are trying to use computers to process content related to discourse coherence. Based on the event chain of the text semantic structure representation, this paper proposes a text semantic structure representation model, on the basis of which, text coherent resources can be used for the task of text semantic analysis. Event chain is a necessary condition for discourse coherence, which can be transformed into a computable event chain analysis problem, and can be further formalized as discourse-oriented partial dependency analysis of sentences.

1 Research Background

"Discourse coherence" has always been one of the hot spots in linguistics. With the development of computer technology and internet technology, the concept of "discourse"has become more and more extensive. The text form gradually entered the field of study of linguists. The distinct features of 21st century text research include: Some corpora marked with part-of-speech and syntax have emerged and have been put into application. Discourse coherence plays an increasingly important role in information processing, sentence ordering in automatic abstracts, target language alignment in machine translation, and language organization in the field of natural language generation[1].

In recent years,discourse coherence analysis models have received more and more attention in the field of natural language processing. More and more researchers have tried to use computers to process the content related to discourse coherence[2].Coherent discourse automatic processing is actually that computers understand and process the natural language. In order to achieve coherence in computer processing, we must first consider the process of coherence in computer processing and its characteristics. The core of this process is that the machine understands coherent texts, and how to formally express the discourse structure so that the machine can process the corresponding discourse structure.

This article attempts to analyze the mechanism of discourse coherence based on event chains, and proposes a new discourse coherence theory and model based on event chains in order to enhance the linguistic assumptions of entity-based discourse coherence models, while using the existing mature sentences and event analysis technology to provide a new approach for the determinating and applying of discourse coherence.

2 Theoretical Basis of Discourse Coherence based on Event Chain

2.1 The feasibility of analyzing and judging textual coherence from the perspective of computational linguistics on events and event chains

Natural language processing refers to the process of dealing with glyphs, speech and semantics in natural language through using computers.With the rapid development of network technology, natural language processing has also entered a booming period. On the one hand, probability theory and data-driven methods have been widely used in natural language processing. On the other hand, the rapid development of network technology has promoted the rapid development of natural language processing. "Event" can represent different concepts in different fields. The general meaning of "event" refers to things having happened, phenomena or extraordinary things, including rituals, competitions, conferences, festivals, artistic performances, emergencies, sports meets, assemblies and commercial operations could be observed[3]. On the basis of which, we have expanded the concept of "events", treating events as images formed by the projection of objective things in the human brain, and semantic systems based on language symbols. In short,

*fanminbbxy@126.com; **2487720824@qq.com

"events" actually happen in the objective world and often involve changes in action, state, or nature. "Event" is realized in the text through the concepts and propositions expressed by the verbs and the arguments that they carry.

Ever since Vendler first proposed event types from the features of verbs, some researchers have often combined events and event structures with the study of syntax and lexical semantics to explore how events map to syntax in.

The characteristics of computational linguistics require that the language theory of coherence must be operable, which is related to the characteristics of computer science. If an essay is provided to the computer to allow it to read or understand whether the text is coherent, it must also give it some knowledge or rules must be given to computer Therefore, operability is the most essential characteristic of computational linguistics[4].

In the field of natural language processing, events have also been widely used due to their good structure. Research on events is usually divided into three levels: (1) event labeling; (2) event relationships, event identification, and event extraction; (3) application techniques of events. These studies already have relatively complete technologies and algorithms, so that events and event structures have certain operability in the process of discourse processing. This is also the main reason why we use events and event structures to analyze discourse coherence.

2.2 How to achieve text coherence based on the event chain model from the perspective of computational models

The study of text coherence based on event chain is reflected and determined by the coherence of the events contained in the text. The coherence of the events is the coherence of the event chain. There are core events and non-core events between events. The aggregation of core events points to the core event chain. When all the core events are connected to each other to form one or more event chains in a certain form. The condition of text coherence is that the text can form coherence if and only if it contains at least one chain of events. In other words, if there are many vocabulary chains in the text without event chains, the text is likely to be incoherent.

For event chains, each sentence in a discourse may have multiple events, then it is possible to construct multiple event chains, and each event chain can be continuous and can be jumpy. This requires research on how to determine the event chain model of a discourse. On the one hand,these choices involve the description of the event chain, and on the other hand, the complexity of the calculation of the event chain[5].

The event chain is discovered by locating the entity chain in the discourse. The event chain that plays a major role is selected to judge whether the discourse is coherent or not and the coherent mode. So, we do not have to calculate the role of each event.We only extract the event chain mainly influencing coherence to calculate, thereby simplifying the calculation process of the text coherence.

3 The Specific Construction of Event Chain Model and the Construction Process of Discourse Coherent Resources

Under normal circumstances, people understand that a text constructs the discourse semantics by extracting major events, which are related to each other to form an event chain just as in figure 1. The underlying concept of the event chain is the lexical chain. Therefore, we decided to use the lexical chain as a clue to determine the event and event chain.

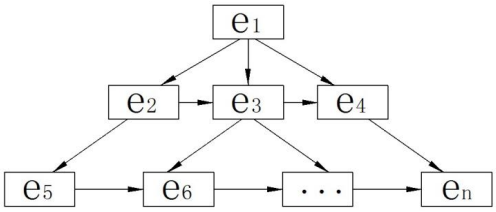

Figure 1. Event hierarchy diagram

3.1 The construction of lexical chains and event chains

A lexical chain is a chain composed of a series of interrelated words. These words are distributed far or near in the text, forming a semantic chain independent of the grammatical structure of the text, but pointing to a common topic. A lexical chain is usually a chain of meaning-related nouns. The event chain in a coherent text is closely related to the lexical chain[6]. Therefore, we can follow the lexical chain clues to build the event chain in the text. The lexical and event chains are intertwined pointing to the central topic of the text. There must be a trunk event chain and multiple branch event chains pointing to the central theme of the discourse. The main event chain runs through the whole text and spans discourses and paragraphs to ensure the overall coherence of the text. The branch event chain crosses sentences and connects adjacent sentences to ensure local coherence.

3.2 Coherent strength calculation

Coherence strength calculation refers to calculating the coherence between events on the entire event chain[7]. After the event chain is constructed, first calculate the coherence strength between each two adjacent events, and then add up all such intensity values, which is the coherence strength of the entire event chain.If the coherence strength between $\triangle e_i$ and $\triangle e_i$ +1 is C_i, the formula for calculating the coherence strength of the entire event chain as shown in figure 2, in the training set, $<e_j , e_k>$is a pair of related event pairs that have been labeled. We have to find such (e_j , e_k)among all the event pairs that have been labeled.Among them, $sim(e_j, \triangle e_i)$ represents the similarity between e_j and $\triangle e_i$. The reason for this is that if there is similarity between the event e_j and $\triangle e_i$, the event e_k is similar to the event $\triangle e_i$ +1.If there exist coherence between event pairs$<e_j , e_k>$,we believe the event $\triangle e_i$ is coherent with the event $\triangle e_i$ +1.

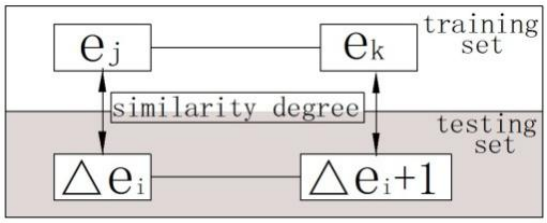

Figure 2. Adjacent event pairs

In other words, the coherence strength between events is calculated as the sum of the similarity between the predicate and the similarity between the arguments (the maximum value among all combinations). As for simWord, it is used to calculate the semantic similarity between words.We turn to Hownet to calculate the semantic similarity between words.

3.3 Evaluation

In this solution, we need to evaluate the lexical chain, event chain, and coherence strength. We use accuracy rate and recall rate as the evaluation criteria. Among them, A is the related lexical pair (or event pair, sentence coherent pair) calculated by our algorithm, and P is the manually labeled related lexical pair (or event pair, sentence coherent pair).

$$Accuracy\ rate = \frac{|A \cap P|}{|A|} \quad (1)$$

$$Recall\ rate = \frac{|A \cap P|}{|P|} \quad (2)$$

Specifically, the evaluation of this study include the following:

Lexical chain: While constructing the event chain, the lexical chain is also determined, which can be evaluated by the correct rate and recall rate of the nodes or edges.

Event: The evaluation of a single event in the event chain can use the usual recall rate and accuracy rate. The accuracy rate includes the accuracy rate for edges and the accuracy rate of the entire event.

Event chain: For coherent discourse, you can determine whether the event chain is correct, and its evaluation includes the accuracy rate and recall rate of all events included in the event chain as a whole.

Coherence: For any text, including the coherent and the incoherent, judgment can be made according to the set threshold, and its evaluation can also be based on accuracy rate and recall rate.

3.4 Construction of Discourse Coherent Resources—The Construction of English Discourse Coherent Resources based on Event Chain

The process of constructing discourse coherent resources is actually the process of constructing a corpus which is of great importance and lays the foundation in the natural language processing. Discourse coherence corpus is a machine-readable electronic resource library composed of real texts that appear naturally and marked with random sampling method according to certain linguistic principles. A good discourse coherence corpus can promote the automatic processing of discourse coherence.The construction of coherent resources in English discourse based on event chain includes the following procedures: corpus selection, corpus preprocessing, lexical chain labeling, event, event chain and event relationship labeling.

3.4.1 Corpus selection

We mainly select English text coherent resources developed by Penn Discouse Treebank as the original corpus. In addition, we have added some recent news releases from major media. The discourse relations are determined by labeling the argument structure, meaning and subordination of discourse connectives and the arguments of discourse connectives.

3.4.2 Corpus pre-processing

Pre-processing includes removing non-standard corpora from the corpus, segmenting and segmenting each news corpus[8]. Eliminating non-standard corpora is done through manual reading. The segmentation and sentence segmentation of a text is accomplished through machines'automatic recognition: the punctuation in the text is used to automatically segment the sentence, and the identifier in the text is used to automatically segment the paragraph.

3.4.3 Determine consistent event labeling criteria

3.4.3.1 Lexical chain annotation

The manual annotation of the lexical chain mainly uses the nouns and nominal phrases in the essays as candidates for constructing the lexical chain, and then inserts the words into the corresponding chain according to the semantic relationships between the words. Create a new chain,if there is not an appropriate chain.

3.4.3.2 Event labeling

Event labeling is to identify events that exist in the text. In theory, the recognition of an event depends on the occurrence of an event trigger word. In most cases, this trigger is based on the appearance of a verb. In order to alleviate the difficulty of computer recognition of events and reduce the obstacles to computing textual coherence, this article adopts the method of manual recognition. When we identify the event, the event will be labeled accordingly.

3.4.3.3 Event chain annotation

Annotating event chains in a discourse is essentially annotating events and event relationships in the discourse. The labeling process can be summarized as: ① determining events in the chapter； ② labeling event relationships；③building event chains. According to our

definition of events (events are things that actually happen in the objective world, usually including changes in actions, states, or properties), we could determine verbs as triggers for events, and use these triggers to determine the corresponding events, and then determine the relationship between these events.

3.4.3.4 Labeling platform

Corpus tagging is a time-consuming and labor-intensive task, which often requires a lot of time and effort. Therefore, developing a convenient and friendly tagging platform can undoubtedly ensure the quality of the corpus tagging and improve the efficiency of the entire tagging process. We adopt the extended Extensible Makeup Language mark the language and label the selected texts. The annotator selects the target item and clicks the corresponding functional keys. The target item is displayed as annotated. The marked content is saved in XML format in the background. The selection of the target item can be a single selection at a time, or multiple selections at a time.

3.4.3.5 Marking quality control

In order to ensure the high-quality labeling of lexical chains, events, and event chains in the discourse, we hired 8 graduate students major in English linguistics to label the lexical chains, events, event chains, and event relationships. Before labeling, we conducted intensive training on labeling standards for 8 labeling personnel. In order to ensure the consistency between each group of taggers, we have designed a software. We labeled 8 taggers with 8 versions of lexical chains, events and event chains for the same document, using V_1, V_2, V_3 ... … V_8, to show that the specific consistency check method is

$$\text{agreement} = n\frac{|V_1 \cap V_2 \cap V_3 \cap \ldots \cap V_n|}{|V_1| + |V_2| + |V_3| + \ldots |V_1|} \quad (3)$$

These consistency values are compared with our preset threshold K (= 0.8) [25]. If it is less than K, we will continue to modify the annotation until the specified threshold is reached; if it is greater than K, the text of these trial bids becomes the event and event chain annotation. The gold standard just as in table1 is used as the standard for subsequent labelling work.

Table 1. Annotation result consistency

	$\lvert V_1 \rvert$	$\lvert V_2 \rvert$	$\lvert V_3 \rvert$	$\lvert V_1 \cap V_2 \cap V_3 \rvert$	agreement
Lexical chain	53	57	51	50	0.909
Event	95	105	88	86	0.887
Event chain	61	67	59	57	0.921
Total	209	229	198	193	0.896

The consistency of lexical chain annotation is 0.909, the consistency of event annotation is 0.887, the consistency of event chain annotation is 0.921, and the overall consistency is 0.896. The stored consistency digits exceed the specified threshold K(=0.8)[25], indicating that the quality of the labelling between the three groups is stable and reliable, and the subsequent labelling work can continue.

3.4.4 Annotate previous results

We randomly selected 20 English news documents from the news of the Pennsylvania Discourse Tree Bank and marked these documents with lexical chains, events, event chains, and event relationships, and plotted the results of the annotations into a chart. See the figure 3 below for details.

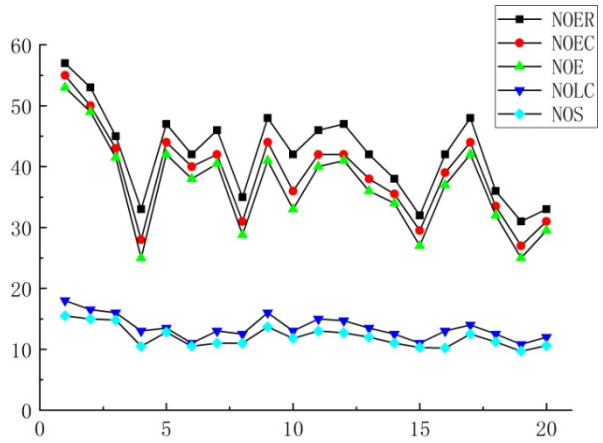

Figure 3. Graph of previous marked results

From the chart, we can see that the number of sentences(NOS) and the number of lexical chains(NOLC) in the consecutive texts are highly correlated, and the number of events(NOE), the number of event chains(NOEC), and the number of event relationships(NOER) are positively correlated. Changes in the number of events, the number of event chains, and the number of event relationships in the discourse have no significant correlation with the number of sentences and the number of vocabulary chains.

4 Application of event chain

4.1 Application in English writing teaching

Discourse coherence is one of the weakest links in the teaching of English writing. Although students' assignments are more accurate in terms of morphology and syntax, if viewed from the overall perspective of discourse, they lack organic cohesion mechanism. Especially in today's network society, resources about writing are extremely rich, but students lack the necessary discourse coherence knowledge. Therefore, it is important to teach students how to write consistently. We can use the principle of event chain to determine whether the text is coherent. The operating procedures used are:① text segmentation and part-of-speech tagging；②discovery of vocabulary chain in the text；③event recognition and extraction in the text；④event chain construction；⑤ text consistency judgment. Between each two adjacent sentence structures in the text, multiple lexical chains can

be created, and these lexical chains can trigger corresponding events. Based on these events, we should determine whether to build an event chain. If they can form an event chain, it can prove that the text is coherent[9]. Otherwise, the text may be incoherent.

4.2 Application in language information processing

Automatic abstraction is the process of using computer to generate abstracts. Automatic digest integrates applied technologies from multiple disciplines such as applied linguistics, computational linguistics, information systems, and artificial intelligence. The most significant advantage of the lexical chain is that it is easy to identify and more convenient to calculate. However, the lexical chain technology has the disadvantage of not being able to take both time efficiency and accuracy into account. Therefore, we try to use the event chain to improve the accuracy rate and recall rate of automatic abstraction, thereby further improving the efficiency of automatic abstraction. At the same time, we plan to build a large set of labelled documents based on the event chain to provide a training set for natural language processing. Through discovering the prominent events in the text, we may extract the sentences where these events are located and arrange them in the order. Create sentences with the strongest chain of events and generate coherent abstracts.

5 Conclusion

This article proposes the concept of “event chain”, updates the theory of discourse coherence, provides a new perspective for the analysis, understanding and generation of discourse, and lays a theory foundation for a comprehensive and multi-faceted investigation of discourse coherence mechanism. This article constructs an English discourse coherence resource based on event chain, establishes a discourse coherence representation model based on event chain, reveals internal and external mechanisms of discourse coherence, and applies the concept of event chain to the evaluation of discourse coherence and the generation of discourse abstracts. In terms of generation, it enriches the way of discourse coherent assessment, and applies the concept of “event chain” in the fields of language teaching, which also has certain guiding significance for the innovation and practice of language teaching theories.

Acknowledgment

This work is supported by the Key Research Project of Social Sciences of Bengbu University under Grant No.2017SK08zd.

References

1. Roman Rodriguez-Aguilar,J.A.Marmolejo-Saucedo. (2019) Structural Dynamics and disruption events in Supply Chains using Fat Tail Distributions. IFAC PapersOnLine., 52: 2686-2691.
2. Arash Azadegan,Ravi Srinivasan,Constantin Blome,Kayhan Tajeddini. (2019) Learning from near-miss events: An organizational learning perspective on supply chain disruption response. International Journal of Production Economics.,38:215-226.
3. Mazzon Giulia,Ajčević Miloš,Cattaruzza Tatiana. (2019) Connected Speech Deficit as an Early Hallmark of CSF-defined Alzheimer's Disease and Correlation with Cerebral Hypoperfusion Pattern. Current Alzheimer research.,16:483-494.
4. Webber B,StoneM,Joshi A, et al.Anaphora and discourse structure. (2003) Computational Lingustics.,29:545-587.
5. Wolf F, Gibson E. (2005) Representing discourse coherence: A corpus-based study. Computational Lingustics.,31:249-288.
6. Htet Myet Lynn, Chang Choi, Pankoo Kim. (2018) An improved method of automatic text summarization for web contents using lexical chain with semantic-related terms.Soft Computing., 22:4013-4023.
7. Lippmann John. (2019) Snorkelling and breath-hold diving fatalities in Australia, 2001 to 2013. Demographics, characteristics and chain of events. Diving and hyperbaric medicine., 49:192-203.
8. Mukherjee Partha,Leroy Gondy,Kauchak David. (2019) Using Lexical Chains to Identify Text Difficulty: A Corpus Statistics and Classification Study. IEEE journal of biomedical and health informatics.,23: 2164-2173.
9. Kong Anthony Pak-Hin,Linnik Anastasia,Law Sam-Po,Shum Waisa Wai-Man. (2018) Measuring discourse coherence in anomic aphasia using Rhetorical Structure Theory. International journal of speech-language pathology.,20:406-421.