英語の中級ノンネイティブ向けの文章における修辞関係のアノテーション

Deng Xinyu　　中村順一

京都大学

概要

グローバル言語である英語は、ノンネイティブに対して、ますます重要になっている。英語の文章では、ノンネイティブにも分かりやすく、よい印象を与えることが重要である。接続詞は、文章において、文脈を理解するのに重要な役割を持つ。英語の中級ノンネイティブ向けの文章における接続詞の位置を研究するために、200,000単語が入っているコーパスを作った。コーパスにおける因果関係を構成する従属接続詞のbecauseとsinceの例文、条件関係を構成する従属接続詞のifとwhenの例文，対比関係を構成する従属接続詞のalthoughとwhileの例文の1072文を収集し、RSTで修辞関係のアノテーションを行った。本論文では、アノテーションの方法を紹介する。

Annotation of discourse relations within the texts whose target audience was intermediate non-native speakers of English

Xinyu Deng　　**Jun-ichi Nakamura**

Kyoto University

Abstract

As an international language, English has become more and more important for non-native speakers. Therefore, authors ought to write English in a way that can be understood quite well by non-native audience. Discourse markers play an important role in keeping the coherence of texts. In order to investigate the position of discourse markers within the texts whose target audience was intermediate non-native speakers, we created a corpus which contains 200,000 words. Using RST, we annotated 1072 examples of three discourse relations, i.e. "reason" relation signaled by discourse marker *because* or *since*, "condition" relation signaled by discourse marker *if* or *when*, and "contrast" relation signaled by *although* or *while*. In this paper, we introduce how to annotate these discourse relations.

1 Introduction

At present, the population of non-native speakers is twice that of native speakers. As a tool for global communication, English has become more and more important in people's daily lives. In order to write English articles which can be understood quite well by non-native audience whose reading ability is lower, it is necessary to explore the texts whose target audience was non-native speakers. Generally, non-native speakers are divided into

three levels: primary (middle school level), intermediate (high school level) and advanced (university level). In this study, we focus on the English texts whose target audience was intermediate non-native speakers.

Since discourse markers play an important role in keeping the coherence of texts, we aim at investigating the position of discourse markers. We collected texts (domain: *natural and pure science*) from high school students' English textbooks published in China and in Japan, and created a corpus TANN (**T**arget **A**udience was intermediate **N**on-**N**ative speakers) which contains 200,000 words. Using Rhetorical Structure Theory (Mann and Thompson, 1988), we annotated three discourse relations, i.e. "reason" relation which is signaled by discourse marker *because* or *since*, "condition" relation which is signaled by discourse marker *if* or *when*, and "contrast" relation[1] which is signaled by *although* or *while*. In this paper, we introduce the first step of the study, i.e. how to annotate the discourse relations. The rest of the paper is arranged as follows. Section 2 introduces the method of annotation. In Section 3, we draw a conclusion.

2 Annotating discourse relations

2.1 Selecting discourse relations

While selecting the discourse relations from TANN, we did not consider the structures such as "not because...but because" and "if...or if...". Lastly, 1072 examples of the discourse relations were selected. The number distribution of the examples is as follows:

Discourse relations	Discourse markers	Number of discourse markers selected
Reason	because	272
	since	46
Condition	if	381
	when	228
Contrast	although	83
	while	62
Total		1072

Table 1: Number distribution of 1072 examples

2.2 Rhetorical Structure Theory (RST)

RST was originally developed for text generation by a team at Information Sciences Institute of University of Southern California. We have two reasons to apply RST to annotation. First, RST is suitable to represent the discourse structure of any genre of texts. Therefore, there is no problem to use RST to annotate the texts whose domain is *natural and pure science*. Second, in RST, the discourse relations initially defined is an open set. That is, the researchers can add or modify relations according to their needs. In this study, we defined 12 discourse relations: background, condition, contrast, elaboration, evaluation, example, list, purpose, reason, restatement, summary and time.

2.3 An example of annotation

Since the aim of our study is to explore the position of discourse markers, we annotate not only the discourse relations signaled by these discourse markers (we call these relations *local relations*) but also the discourse relations which contain the *local relations* (we call

[1]In this study, "contrast" relation refers to both "contrast" relation and "concession" relation.

these relations *whole relations*). For example, in the text "When exposed to white light, a white object looks white because it reflects all colours." (the RST analysis of this text is shown schematically in Figure 1), discourse marker *because* signals "reason" relation between the main clause "a white object looks white" (i.e. nucleus) and the subordinate clause "it reflects all colours" (i.e. satellite). The "reason" relation is the *local relation.* On the other hand, discourse marker *when* signals "condition" relation between the sentence "a white object looks white because it reflects all colours" (i.e. nucleus) and the non-finite clause "exposed to white light" (i.e. satellite). The "condition" relation is the *whole relation.*

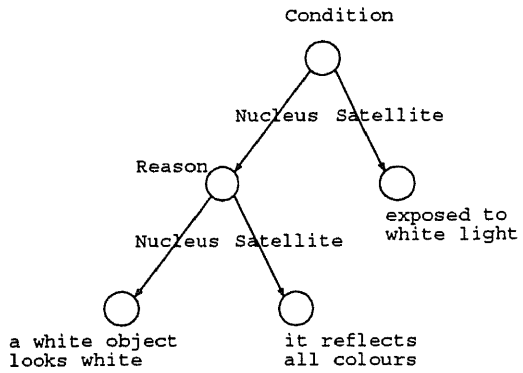

Figure 1: An example of analysing the structure of text by RST

Annotating the text shown in Figure 1 can be individed into the following two steps:

1. Annotating the boundary of the "reason" relation which is signaled by discourse marker *because* in round brackets, and then labeling its nucleus (N) and the satellite (S), i.e.

   When exposed to white light, (a white object looks white) -N–reason–S- (because it reflects all colours).

2. Annotating the boundary of the "condition" relation signaled by *when* in angle brackets, and then labeling its nucleus (N) and satellite (S), i.e.

   <When exposed to white light,> -S–condition–N- <(a white object looks white) -N–reason–S- (because it reflects all colours)>.

2.4 Training the coders

In order to make the annotation results precise and reliable, we wrote a reference manual for tagging the discourse relations and trained two independent coders before annotating all discourse relations selected. The two coders were asked to annotate a small test corpus[2] for three times, which lasted two weeks respectively.

Training the coders can be divided into the following three steps:

1. The two coders were asked to read the reference manual and annotate the relations within the test corpus according to their understanding of the manual. The rate of

[2]The small test corpus contains the first 120 examples of the "reason" relations signaled by *because*.

agreement between the two coders[3] was 45.9% (the second column of Table 2). Then we analysed the problems that caused the disagreement of the two coders and revised the manual.

2. The two coders were asked to annotate the relations within the test corpus again according to the revised manual. The rate of agreement between the two coders became 64.9% (the third column of Table 2). We analysed the problems that caused the disagreement of the two coders and revised the manual again.

3. The two coders were asked to annotate the test corpus according to the newly revised manual. Since the rate of agreement between the two coders were higher than (or equal to) 95% on the three aspects (the forth column of Table 2), we stopped training the coders. The newly revised manual would be used as the reference manual (see Appendix) later.

		1st time	2nd time	3rd time
Period of time during annotation		two weeks	two weeks	two weeks
Rate of agreement	Boundary	87.5%	92.5%	95.8%
	Discourse relation	69.2%	82.5%	95.0%
	Nucleus and satellite	75.8%	85.0%	96.7%
Total		45.9%	64.9%	88.0%

Table 2: The rate of agreement between the two coders for annotating the test corpus

As shown in Table 2, before training, the agreement of the two coders was 45.9% (the second column). However, after training, the agreement of the two coders became 88.0% (the forth column), which was 42.1% higher than that before training. This shows that training coders can improve the rate of agreement between the two coders.

2.5 Annotating discourse relations

The two trained coders took part in annotation. Of the two coders, one was main coder, and another was reliability coder. The main coder annotated the 1072 examples selected. We will use the annotation results of the main coder to do experiments. The reliability coder annotated the first 30 examples of the discourse relations signaled by the six discourse markers mentioned above respectively. That is, the reliability coder annotated 180 examples of discourse relations. The annotation lasted three months.

In order to assess the reliability of annotation, we compared the results of the 180 examples annotated by the reliability coder with those annotated by the main coder from three aspects (i.e. the boundary, discourse relation, nucleus and satellite of the *whole relation*). Table 3 shows that the rate of agreement between the reliability colder and the main coder was 82.3%. This result was higher than that mentioned in (Moser and Moore, 1995). We think that the reference manual of annotation was very helpful for the coders, because the precise definition of each relation avoided misunderstanding. Furthermore, the two trained coders had linguistic background, so they could quite grasp the meaning of the manual.

[3]We follow Moser and Moore's approach (1995) to assess the reliability of annotation. We assessed the agreement of annotation results of the *whole relation* from three aspects, i.e. boundary, discourse relation, nucleus and satellite.

Whole relation	Rate of agreement
Boundary	93.9%
Discourse relation	91.7%
Nucleus and satellite	95.6%
Total	82.3%

Table 3: The rate of agreement between the reliability coder and the main coder

3 Conclusion

This paper introduces the method of annotating 1072 examples of three discourse relations within corpus TANN by the framework of RST. These discourse relations are: "reason" relation which is signaled by *because* and *since*, "condition" relation which is signaled by *if* and *when*, "contrast" relation which is signaled by *although* and *while*. In order to make the annotation results reliable, we wrote a reference manual and trained two independent coders before annotation.

We assessed the reliability of annotation by analysing the rate of agreement between the reliability coder and the main coder from three aspects, i.e. boundary, discourse relation, nuclear and satellite of the *whole relation*. The analysis results showed that the rate of agreement between the two coders were higher than 90% on the three aspects respectively. The annotation results will be used to investigate the position of the discourse markers (i.e. *because*, *since*, *if*, *when*, *although* and *while*) within the texts whose target audience was intermediate non-native speakers of English.

Appendix

Reference Manual for discourse annotation

1. Introduction

This reference manual presents a guideline to annotate discourse relations using the framework of Rhetorical Structure Theory (RST). RST was originally developed for text generation by a team at Information Sciences Institute of University of Southern California. It points out that the discourse relations that hold between text spans make text coherent (more details about RST could be found in (Mann and Thompson, 1988)).

According to RST, each text span is categorized as a nucleus or a satellite. A mononuclear relation (e.g. "condition" relation and "reason" relation) contains a nuclear and a satellite. A nucleus represents the essential piece of information in the relation, while a satellite indicates supporting or background information. Compared with satellites, nuclei play a crucial role in keeping the coherence of a text. A multinuclear relation (e.g. "list" relation) contains two nuclei of equal importance in the discourse relation. We defined 12 discourse relations which are introduced in Section 2.

2. Definitions of discourse relations

2.1 Background: mononuclear

Definition: in a "background" relation, the situation presented in the satellite provides the

context in which the situation presented in the nucleus should be interpreted. However, the satellite is not the reason of the situation presented in the nucleus. The situation presented in the satellite is objective, and the reader/writer intentions are irrelevant in determining whether such a relation holds.

Example: (In 1962 the book titled 'Silent spring' was published and caused a greater stir than anyone had ever imagined.) -N–background–S- (This monumental work in ecology made people aware of the dangers of chamical insecticides and changed the course of our history.)

2.2 Condition: mononuclear

Definition: in a "condition" relation, the truth of the proposition of the nucleus is a consequence of the fulfilment of the condition in the satellite. Sometimes, a "condition" relation is signaled by a discourse marker, such as *if* and *when.*

Examples:

1. (If you do not go to bed early,) -S–condition–N- (you cannot have enough sleep.)
2. (Some birds will fly away to the south) -N–condition–S- (when the weather turns cold.)

2.3 Contrast: mononuclear

Definition: in a "contrast" relation, the situation presented in the nucleus is contrary to expectation in the light of the information presented in the satellite, or the situation presented in the nucleus comes in contrast with the situation presented in the satellite. Sometimes, a "contrast" relation is signaled by a discourse marker, such as *but* and *while.*

Examples:

1. (The sun heats the entire earth – the land, oceans, and air.) -S–contrast–N- (But these three materials do not all heat up at the same rate.)
2. (Small winds can cause ripples,) -N–contrast–S- (while strong winds create large hurricane waves.)

2.4 Elaboration: mononuclear

Definition: in an "elaboration" relation, the satellite gives additional information (or detail) about the situation or a part of the situation presented in nucleus.

Example: (Vitamins were unknown until the beginning of the twentieth century.) -N–elaboration–S- (Between 1915 and 1945, over 50 different substances were discovered in food and many were found to be substances that body could make for itself, therefore the

number was reduced to 15 essential vitamins – vitamin A, Vitamin B complex, Viatmin C, Vitamin D, Vitamin E and Vitamin K.)

2.5 Evaluation: mononuclear

Definition: in an "evaluation" relation, the satellite assesses the situation presented in the nucleus. An evaluation can be an appraisal, estimation, rating. The evluation can be the viewpoint of the writer or another agent in the text.

Example: (Some people don't like canned or frozen vegetables, because they think fresh vegetables cooked at home are always better.) -N–evaluation–S- (This is wrong.)

2.6 Example: mononuclear

Definition: in an "example" relation, the satellite gives an example to the information or situation presented in the nucleus. Sometimes, an "example" relation is signaled by a discourse marker, such as *for example* and *for instance.*

Examples:

1. (In tropical areas, houses are sometimes made from the plants that grow there.) -N–example–S- (For example, houses in Africa or Asia may be made out of bamboo.)

2. (Most of the world's highest mountains were formed quite recently in earth's history.) -N–example–S- (For instance, the Himalayan mountains have built up within the last 40 million years and they are still growing even today.)

2.7 List: multinuclear

Definition: a "list" relation is a multinuclear relation whose elements can be listed, but which are not in a contrast relation.

Example: The answer lies in two facts. (The first is that it has strored supplies of fat in its body during the summer and autumn.) -N–list–N- (The second is connected with the main use the body makes of food – to supply the energy for movement.)

2.8 Purpose: mononuclear

Definition: in a "purpose" relation, the situation presented in the satellite is only putative, i.e. it is yet to be achieved. Most of it can be parapharased as "nucleus in order to satellite".

Example: (In order to answer these questions,) -S–purpose–N- (NASA alunched a spaceship, the Mars Pathfinder, in December, 1996.)

2.9 Reason: mononuclear

Definition: in a "reason" relation, the situation presented in the satellite is the reason of the situation presented in the nucleus. Sometimes, a "reason" relation is signaled by a discourse marker, such as *because* and *so*.

Examples:

1. (Elephants often coat their skin with mud,) -N–reason–S- (because it keeps them cool and protects them from insects.)
2. (These buildings were over 60 years old,) -S–reason–N- (so they were not strong enough.)

2.10 Restatement: mononuclear

Definition: in a "restatement" relation, the satellite reiterates the information presented in the nucleus, typically with slightly different wording. It does not add to or interpret the information.

Example: (Save the earth.) -N–restatement–S- (Save our planet.)

2.11 Summary: mononuclear

Definition: in a "summary" relation, the satellite summarizes the information presented in the nucleus.

Example: (After thousands of years of selecting, or choosing the biggest seeds, farmers ended up with what we know today as wheat.) -N–summary–S- (It came from nothing more than ordinary grass.)

2.12 Time: mononuclear

Definition: in a "time" relation, the situation presented in the nucleus occurs after (or before, or at the same time) the situation presented in the satellite.

Examples:

1. (After Asian elephants have been captured,) -S–time–N- (they are easily trained.)
2. (We have a long way to go) -N–time–S- (before people live on the moon.)
3. (While Armstrong was landing on the moon's surface,) -S–time–N- (Eagle almost ran out of fuel.)

References

Mann, W. and Thompson, S. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text, 8(3)*, 243-281.

Moser, M. and Moore, J. (1995). Investigating cue selection and placement in tutorial discourse. In *Proceedings of the 33rd ACL.*