社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－NL－170（14）
2005／11／21

ノンネイティブユーザを対象にした接続詞の用法に関する研究

Xinyu Deng　　中村順一

京都大学

概要

世界中で、ノンネイティブの人口はネイティブの倍になっている。英語はノンネイティブに対しても、大切な言語である。いままで、文章生成という分野の研究のほとんどは、英語のネイティブを対象にしている。本研究では、中級レベルのノンネイティブを対象にして、接続詞の用法を分析した。とりわけ、接続詞 *because* について、between-text-span punctuation に影響のある要素を調べた。Quinlan による決定木生成アルゴリズムの C4.5 を利用して、between-text-span punctuation のクラス判別モデルを作成した。その上で、接続詞 *but*、*for example* 及び *if* の用法も分析した。研究の結果は開発されている中級レベルノンネイティブ向けの文章生成システム (The SILK System) に応用する。

Investigating the cue usage for non-native users

Xinyu Deng　　Jun-ichi Nakamura

Kyoto University

Abstract

At present, the population of non-native speakers is twice that of native speakers. It is necessary to explore the text generation strategies for non-native users. This study investigates the features that affect the between-text-span punctuation of *because* while it signals "explanation" relation for non-native speakers. A machine learning program – C4.5 was applied to induce the classification models of the punctuation. In addition, the usage of *but*, *for example* and *if* were also investigated. The experiment results will apply to the SILK (Generation *S*system for *I*ntermediate *L*evel non-native spea*K*ers on discourse level) system which we are developing.

1 Introduction

As an international language, English has become more and more important for non-native speakers. However, almost all English documents are written for the native speakers. To some degree, some documents can not be understood quite well by non-native speakers. Until now, little research has been done for non-native users in the field of Natural Language Generation (NLG). In this study, we concentrate on exploring the cue usage on discourse level. Our aim is to find the decision-making mechanisms of text generation for non-native users.

From the viewpoint of NLG, the following three texts have the same abstract structure, though the differences among them are apparent. E.g., cue placement (where the cue should be placed in the text) is different. In text 1, cue phrase *because* occurs in the first span, while in text 2 and 3, cue phrase *because* occurs in the second span. Moreover, between-text-span punctuation is different. In text 2, the between-text-span punctuation of *because* is a comma. But in text 3, the punctuation is a space, i.e., no punctuation is used. (Deng and Nakamura, 2005) investigated the feature affecting the placement of *because* while it signals "explanation" relations. By a machine learning program – C4.5, this study explore the feature(s) affecting the between-text-span punctuation when the word *because* occurs in the second span.

1. Global warming will be a major threat to the whole world over the next century. ***But because*** it will take many years for our actions to produce a significant effect, the problem needs attention now.

2. Global warming will be a major threat to the whole world over the next century. ***But*** the problem needs attention now, ***because*** it will take many years for our actions to produce a significant effect.

3. Global warming will be a major threat to the whole world over the next century. ***But*** the problem needs attention now ***because*** it will take many years for our actions to produce a significant effect.

Generally, non-native speakers are divided into three levels: primary (middle school student level), intermediate (high school student level) and advanced (university student level). The users of this study are assumed to be at intermediate level. In this research, we confine ourselves to the texts whose domain is *natural and pure science*. The analysis results will be applied to the SILK system. The rest of the paper is arranged as follows. Section 2 describes related work. Section 3 demonstrates how to create and annotate a corpus. In Section 4, a machine learning program – C4.5 is introduced. Section 5 shows the experiment results. Section 6 introduces the usage of four cue phrases. Section 7 draws a conclusion.

2 Related work

Almost all researches on cue phrases have been done for native speakers. (Elhadad and McKeown, 1990) explored the problem on cue selection. They presented a model that distinguishes a small set of similar cue phrases. (Moser and Moore, 1995a) put forward a method to identify the features that predict cue selection and placement. (Eugenio and Moore and Paolucci, 1997) used C4.5 to predict cue occurrence and placement. (Williams, 2004) measured the differences of reading speed (especially cue phrases) between good readers and bad readers, by which they inferred how discourse level choice (e.g., cue selection) makes the difference for the two kinds of readers.

3 Corpus

In order to explore the cue usage for non-native users, we created and annotated a corpus.

3.1 Creating corpus

CNNSE (Corpus for Intermediate Level Non-Native Speakers of English), whose size is 200,000 words, was created by the first author. According to the Flesch Reading Ease scale, the readability of CNNSE is 68.7 (easy). We extracted English texts (written or rewritten by native speakers) from the books published in China and in Japan. The target audiences of these books were high school students in the two countries. The domain of the selected texts is *natural and pure science.*

3.2 Annotating corpus

Cue phrase *because*, which signals "explanation" relation, was annotated according to RST (Mann and Thompson, 1988). Annotation includes two stages: first, we allowed two coders to choose *because* using (Hirschberg and Litman, 1993)'s 3-way classification. The word *because* could signal not only "explanation" relation, but other relations. On the other hand, we do not consider some structures, e.g., "not because ... but because". Thus, *because* could be judged as "explanation", "other", or "not considered". If both coders classified *because* as "explanation", this discourse was selected. Lastly, 124 *because* were selected, in which 88 *because* occur in the second span. At the second stage, two coders annotated the boundary of nucleus and satellite of each discourse selected. Moreover, a selected discourse could be a span (nucleus or satellite) of another one (we call it embedding structure). The coders labeled the discourse relation of the embedding structure and determined the boundary of its nucleus and satellite. An example is shown as follows.

[Global warming will be a major threat to the whole world over the next century.]–S– contrast –N–[But [because it will take many years for our actions to produce a significant effect,]–S– explanation –N–[the problem needs attention now.]] (From CNNSE)

In order to assess reliability of annotation, we followed (Moser and Moore, 1995b)'s approach to compare the disagreements of results annotated by two independent coders from three aspects. First, the boundary of nucleus and satellite of the relation signaled by *because*; the disagreements occurred 5 times (96.0% agreement). Second, the discourse relation of embedding structure; the disagreements occurred 9 times (92.7% agreement). Third, the boundary of nucleus and satellite of the embedding structure; the disagreements occurred 6 times (95.2% agreement). That is, the agreement of the two coders is 83.9%.

4 Machine learning program – C4.5

C4.5 is a set of computer programs that examine numerous recorded classifications and construct a model inductively by generalizing from specific examples (Quinlan, 1993). Its main function is identifying and analysing patterns in amount of data. We applied C4.5 to induce the classification models of between-text-span punctuation when *because* occurs in the second span.

4.1 Evaluation method

The results of C4.5 are learned classification models from the training sets. The error rates of the learned models are estimated by *cross-validation* (Weiss and Kulikowski, 1991), which is widely applied to evaluating decision trees, especially whose dataset is relatively small. In this study, data for learning was randomly divided into 10 test sets. The program was run for 10 times, each run used 9 test sets as the training set and the remaining one as the test set. The error rate of a tree obtained by using the whole dataset for training was then assumed to be the average error rate on the test set over the 10 runs. The advantage of this method is that all data are eventually used for testing, and almost all examples are used in any given training run (Litman, 1996). The method of determining whether two error rates are significantly different is by computing and comparing the 95% confidence intervals for the two error rates. If the upper bound of the 95% confidence interval for error rate $\varepsilon 1$ is lower than the lower bound of the 95% confidence interval for $\varepsilon 2$, then the difference between $\varepsilon 1$ and $\varepsilon 2$ is considered to be significant.

4.2 Features

We classified features into two groups: sentence features and embedding structure features. Sentence features are concerned with the information of relations signaled by *because*. Embedding structure features reflect the information of the embedding structures that contain relations signaled by *because*. Features used in the experiments are as follows:

- Sentence features
  - Nt. Tense of nucleus: past, present, future.
  - St. Tense of satellite: past, present, future.
  - Nv. Voice of nucleus: active, passive.
  - Sv. Voice of satellite: active, passive.
  - Ng. Length of nucleus (in words): integer.
  - Sg. Length of satellite (in words): integer.
  - Ns. Structure of nucleus: simple, other.
  - Ss. Structure of satellite: simple, other.
- Embedding structure features
  - R. Discourse relation of embedding structure: attribution, background, cause, etc.
  - C. Signaled by cue or not: yes, no.
  - N-S. Role of the relation signaled by *because*: nucleus, satellite.
  - P. Position of relation signaled by *because*: first span, second span.
  - Bg. Length of the span containing the relation signaled by *because*: integer.
  - Og. Length of the span not containing the relation signaled by *because*: integer.
  - Bs. Structure of the span containing the relation signaled by *because*: complex sentence, other.
  - Os. Structure of the span not containing the relation signaled by *because*: simple sentence, other.

5 Experiment results on between-text-span punctuation

First we introduce a concept – baseline, which can be obtained by choosing the majority class. E.g., in this research, when the word *because* occurs in the second span, 73.9% (65/88) between-text-span punctuation is space (no punctuation). That is, if no punctuation is used, one would be wrong 26.1% of the times. So 26.1% is the error rate of the baseline model that is used in this experiment. The experiment had four sets: Experiment Set 1 were run for examining the best individual feature whose predictive power was better than the baseline; Experiment Set 2, 3 and 4 were run for classifying the between-text-span punctuation when *because* occurs in the second span.

In Experiment Set 1, we ran the experiment 16 times using each feature mentioned above. We found that only feature Bg has predictive power, for the 95% confidence interval of its error rate was 13.5 ± 2.4, whose upper bound for error rate (15.9%) was much lower than the baseline (26.1%). That is, when the length of the span containing the relation signaled by *because* is less than (or equal to) 21 words, no punctuation is used, otherwise, a comma is used.

Experiment Set 2 (Table 1) had four subsets. Each experiment was run only using sentence features. In subset 1, all eight sentence features were used. Then we ran three other experiments using a combination of different sentence features. For all the four subsets, the upper bound of the 95% confidence interval for error rate were higher than the baseline (26.1%). So the learned models were not good ones.

Table 1: Feature sets and results of Experiment Set 2

	Nt	St	Nv	Sv	Ng	Sg	Ns	Ss	R	C	N-S	P	Bg	Og	Bs	Os	Result
1	x	x	x	x	x	x	x	x									27.1 ± 5.3
2	x	x	x	x	x	x											24.7 ± 4.5
3	x	x	x	x			x	x									26.9 ± 5.4
4					x	x	x	x									28.0 ± 5.6

Experiment Set 3 (Table 2) had four subsets, in which both sentence features and embedding structure features were used. In subset 1, experiment was run using all sixteen features. Experiment result showed that the upper bound of the 95% confidence interval for error rate (23.6%) is lower than the baseline (26.1%). It proves that feature Bg could improve the accuracy of the learned models. However, the learned model was not the best one. Then we tried three other feature combinations. Experiment results showed that the three learned models were not good ones.

Table 2: Feature sets and results of Experiment Set 3

	Nt	St	Nv	Sv	Ng	Sg	Ns	Ss	R	C	N-S	P	Bg	Og	Bs	Os	Result
1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	22.3 ± 1.3
2							x	x				x			x	x	25.9 ± 5.1
3	x	x	x	x			x	x	x	x	x	x			x	x	28.3 ± 5.4
4					x	x							x	x			32.5 ± 3.1

Experiment Set 4 (Table 3) had four subsets, in which only embedding structure features were used. In subset 1, the experiment was run using all the eight embedding structure features. The upper bound of the 95% confidence interval for error rate (23.2%) of the learned model was lower than the baseline. It proves that Bg is the feature that affects the accuracy of learned models again. In subset 2, 3 and 4, we ran the experiment by deleting features representing span length (Bg and Og), feature R, and features representing structure (Bs and Os) respectively. However, no good model was obtained.

Table 3: Feature sets and results of Experiment Set 4

	Nt	St	Nv	Sv	Ng	Sg	Ns	Ss	R	C	N-S	P	Bg	Og	Bs	Os	Result
1									x	x	x	x	x	x	x	x	20.9 ± 2.3
2									x	x	x	x			x	x	29.1 ± 5.9
3										x	x	x	x	x	x	x	31.4 ± 4.1
4									x	x	x	x	x	x			30.3 ± 4.5

6 Usage of cue phrases

Besides cue phrase *because*, we investigated the usage of three other cue phrases (*but*, *for example*, and *if*) by the same method. The cue usage of the four cue phrases are summarized in Table 4:

Table4: Usage of cue phrases

	Placement	Between-text-span punctuation
because	first span (if R is “contrast”, “example”, or “explanation”)	comma
	second span (if R is not “contrast”, “example”, or “explanation”)	no punctuation (if $\mathsf{Bg} \leq 21$) comma (if $\mathsf{Bg} > 21$)
but	second span	comma (if $\mathsf{Bg} \leq 29$) full stop (if $\mathsf{Bg} > 29$)
for example	second span	full stop
if	first span	comma

7 Conclusion

This study investigates the cue usage for non-native users. We introduce a method to induce the best classification model of between-text-span punctuation when cue phrase *because* occurs in the second span. The experiment results showed that length of the span which contains the relation signaled by *because* is the most powerful feature. The heuristics obtained from machine learning experiments can be applied to NLG systems.

References

Xinyu Deng and Jun-ichi Nakamura. 2005. *Investigating the features that affect cue usage of non-native speakers of English.* Poster in the Proceeding of The Second International Joint Conference on Natural Language Processing.

Barbara Eugenio and Johanna Moore and Massimo Paolucci. 1997. *Learning Features that Predict Cue Usage.* Proceedings of the 35th Conference of the Association for Computational Linguistics.

William Mann and Sandra Thompson. 1988. *Rhetorical structure theory: Toward a functional theory of text organization.* Text, 8(3).

Diane Litman. 1996. *Cue Phrase Classification Using Machine Learning.* Journal of Artificial Intelligence Research, Vol.5, 53-94.

Julia Hirschberg and Diane Litman. 1993. *Empirical studies on the disambiguation of cue phrases.* Computational Linguistics, 19(3) 501–530.

Megan Moser and Johanna Moore. 1995a. *Using discourse analysis and automatic text generation to study discourse cue usage.* AAAI Spring Symposium Series: Empirical Methods in Discourse Interpretation and Generation, 92-98.

Megan Moser and Johanna Moore. 1995b. *Investigating cue selection and placement in tutorial discourse.* Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics.

Michael Elhadad and Kathleen McKeown. 1990. *Generating connectives.* Proceedings of the 12th International Conference on Computational Linguistics.

Sandra Williams. 2004. *Natural language generation (NLG) of discourse relations for different reading levels.* Ph.D. Thesis, University of Aberdeen.

Ross Quinlan. 1993. *C4.5: Programs for Machine Learning.* Morgan Kaufmann.

Sholom Weiss and Casimir Kulikowski. 1991. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems.* San Mateo, CA: Morgan Kaufmann.