The Fifteenth International Symposium on Artificial Life and Robotics 2010 (AROB 15th '10),
B-Con Plaza, Beppu,Oita, Japan, February 4-6, 2010

A Corpora-based Detection of Stylistic Inconsistencies of Text in the Targeted Subgenre

Kiyota Hashimoto*, Kazuhiro Takeuchi**, Hideaki Ando**

Osaka Prefecture Univ.: 1-1, Gakuen-cho, Naka-ku, Sakai, 599-8531, Japan*
*Osaka Electro-Communication Univ**., 18-8, Hatsu-machi, Neyagawa, 572-8530, Japan*
(Tel : 81-72-252-1161; Fax : 81-72-254-99441)
(hash@lc.osakafu-u.ac.jp)

Abstract: The authors are currently on the way to develop a couple of educational applications for learners to improve their utterance/writing skills with a particular reference to stylistic coherence: a visual aid for teachers and learners to detect style inconsistencies with advice on improvements, an evaluation aid for teachers to grade learners' writings. In this paper, as a foundational data accumulation and analyses for them, we propose a method using multi-corpora comparison to correctly extract expressions not suited to a particular subgenre intended.

Keywords: computational stylistics, writing support, expression classification, visualization, educational tool

I. INTRODUCTION

Any kind of text expressions, from human utterances and writings to robotic utterances and text generations, require stylistic consistency suited to the targeted genres and the author's communicative purposes. Particularly in the educational field, automatic error detection and proof-reading tools are effective for students to learn how to write good essays. However, most of these tools including grammar and style checkers implemented in commercial word processing software ignore genre/subgenre differences and thus the user do not know whether the expressions that are pointed out as bad or wrong are indeed bad or wrong for the targeted genre/subgenre. So alternative method for detecting undesirable expressions for the targeted genre/subgenre is required.

II. GENRE AND STYLISTIC CONSISTENCY

Any text is classified into a genre. A genre is a set of criteria for a category of text, usually according to its topic and its way of publication. Thus, a text may be categorized as in the genre of biology according to its topic, or as a newspaper article because it is published in a newspaper. A genre is further divided into subgenres, part of which is sometimes regarded as linguistic register, mainly according to narrative aspects. How to narrate, or style, is highly related to the targeted audience and the author's communicative purpose of the text, and stylistic consistency is required in a text, though deliberate inconsistencies bring extra literary effects.

Stylistic consistency constitutes the use of suitable words, grammatical expressions, syntactic word orders and complexities, average sentence length, and information flow, though not limited to them. Several studies have tackled with style; particularly it has long been pointed out that basic stylistic consistency is held by the restrictive use of functional expressions particularly in the case of Japanese and other languages that have a rich variety of stylistic grammar forms.

a. Kore-wa hon-desu.
this book be-pres.
This is a book.
b. Kore-wa hon-da.

Both 'desu' and 'da' are auxiliary copular verbs and the difference is up to politeness, which in turn should be determined according to the targeted audience and the author's communicative purpose. So the mixed use of 'desu' and 'wa' causes undesirable stylistic inconsistency and thus should be avoided.

However, finer-grained observations of various kinds of human texts have found that a text genre traditionally considered to hold one single style should be decomposed into several different styles according to their subtle differences of the targeted audience and the author's communicative purposes. For example, we can easily distinguish a newspaper article and a newspaper editorial, or a textbook for graduate students and one for undergraduate or high school students, not just in terms

of their contents but in terms of their style, though few of us can always make clear our criteria for this kind of distinction. Style has been studied mainly in the field of linguistics, literature, and education (for example, [1], [2]) but most of them are based on subjective, aesthetic judgments, and the finer-grained distinction of subgenres requires more objective, corpora-based analysis. From this viewpoint, in order to develop educational tools for better utterances and texts, it is necessary to first develop computational tools for detecting and evaluating stylistic inconsistencies.

III. STYLISTIC FEATURE EXTRACTION AND STYLISTIC VISUALIZATION

Stylistic consistency does not only depend on the inner consistency in the text, but also on the appropriate choice of style for the textual purpose. In other words, the targeted audience and the author's communicative purpose determine the desirable style; then the author, with his/her limited reading experiences, tries to use as many appropriate stylistic features as possible and tries not to use inappropriate stylistic features. Very often the author, particularly the learning one, makes mistakes on the choice of appropriate expressions partly because the stylistic distinction is subtle and often unconscious, partly because the author unconsciously relies on his/her judgments that comes from the most accustomed style, informal speech style, and partly because every one of us speaks and writes a number of different texts each of which has its own style and tends to rely on his/her intuition which only tells them that a particular stylistic feature may not suit the purpose.

With these in mind, we first construct a set of textual corpora that consists of two or more subgenres and extract stylistic features of each subgenre. This set roughly corresponds to our reading experiences but the larger size is naturally expected to contribute to a better detection of stylistic features. For the first approximation, let us consider that stylistic features of a subgenre are based on the preferable and unused expressions for the subgenre. The definition of 'expression' may vary, but we should note that a large number of words, including misspelled ones, are found unregistered in the dictionaries that any taggers use, that our current target language, Japanese, has still some serious problems in tokenization, and, most importantly, that the mere frequency of a word or a phrase without regard to its context does not reflect the precise tendency of the use of the word in a subgenre because a word usually has more than one meaning and usage, each of which are differently preferred or avoided in different subgenres. So here, we tentatively define an expression as an n-gram character string. The n-gram character extraction is a simple way of extracting expressions, but successfully contains contextual information when n is large enough, though too large n naturally causes too many low frequencies. The determination of the proper (range of) n is a heuristic issue, and we first adopt the range of n as two to four, partly because most of the Japanese words consist of one or two characters.

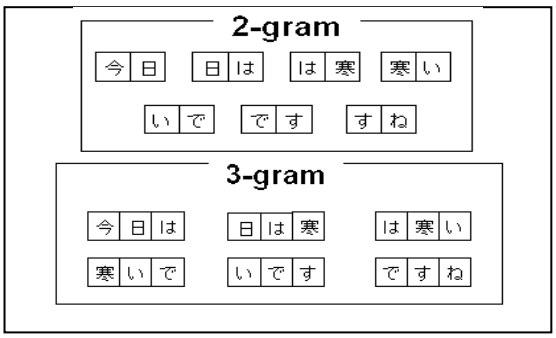

Fig.1. n-gram sample expressions in a Japanese equivalent of 'It is cold today'

As sample contrastive data, we use a dumped file of Japanese Wikipedia [3] and a sample of 2ch BBS (http://www.2ch.net/).

Data	Number of characters
Wikipedia	161,223,892
2ch	108,031,243

Table 1. Data set

The frequency of each expression in Wikipedia and 2ch classifies the expressions roughly in three classes: (a) expressions frequently used in Wikipedia but not frequently used in 2ch, (b) expression frequently used in both data, and (c) expressions not frequently used in Wikipedia but frequently used in 2ch. This classification means that the class (b) consists of rather neutral expressions, but the class (a) and (c) are to be considered to be stylistic features for Wikipedia and 2ch, respectively. The scatter diagram of 2- to 4-gram expressions is shown in Fig.2.

With this kind of scatter diagram, each expression, or stylistic feature, can be graded according to its

distance from the catercorner([4]). Let a function gr(*e*) be defined as follows:

For a given expression *e*, the grade function gr (*e*) of Wikipedia-preferred expressions is

gr (*e*) = c dis (*e*)

where dis (*e*) is the distance from the catercorner,

and c = 1 if *e* is plotted below the catercorner

c = 0 if *e* is plotted on the catercorner

c = -1 if *e* is plotted above the catercorner.

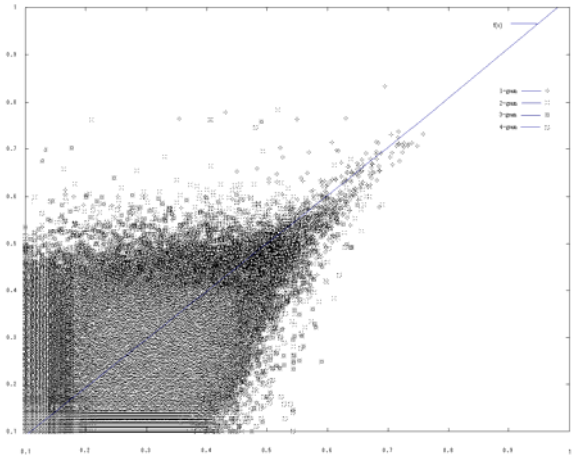

Fig.2 Scatter Diagram of 2- to 4-gram expression (X: frequencies in Wikipedia, Y: frequencies in 2ch)

Then, a given text *T* can be analyzed using gr (*e*) for every expression in *T* in terms of the resemblance with Wikipedia or 2ch.

We made an expression database $\Sigma(e)$ consisting of gr (*e*) of all the expressions in the Wikipedia and 2ch corpora we used. Then every expression that appears in a given text *T* is graded using $\Sigma(e)$, with which we obtain a sequence of gr (*e*) of *T*. A sample visualization of a student' essay is shown in Fig. 3.

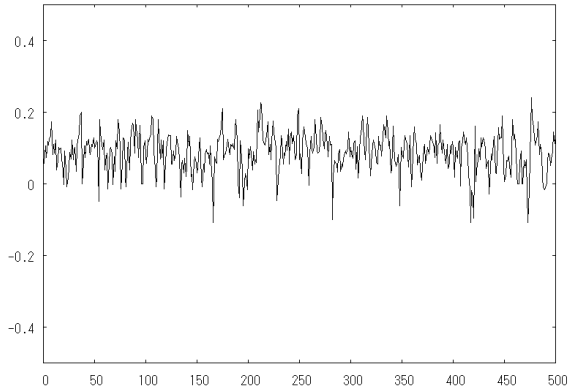

Fig. 3 Resemblance Visualization

As shown in Fig. 3, the overall tendency of the student's essay used for this analysis resembles more to Wikipedia than to 2ch, but still it contains lots of expressions that are avoided in Wikipedia or preferred in 2ch. In other words, if we have a model set of corpora, one of which represents the targeted subgenre, any text can be graded and visualized with the function gr (*e*), and this type of visualization, though more improvements are necessary for practical use, can be used to visually point out expressions that may be avoided for the targeted subgenre.

IV. IMPROVEMENTS

In the previous section, we proposed a rather simple approach to evaluate a text in term of its appropriate style for the targeted subgenre by comparing its expressions with a model set of corpora, and visualized the result. In order to apply this method to an educational tool, at least two problems are to be solved. First, the method only uses the frequency of expression to judge whether a given expression is preferable or not for the targeted subgenre, but there should be different reasons for each expression being judged unfavorable: some may contain spelling errors, some grammatical errors, and others undesirable choice of words or phrases. Second, the visualization like Figure 3 is redundant for educational purposes, since what is important is to point out unfavorable expression for the targeted subgenre, and the grade differences among preferable expressions make little sense because the value of gr (*e*) near zero means that the expression tends to be used neutrally among corpora.

For dealing with the first problem, an additional method detecting serious errors is required. As for the determination of the serious errors, we manually observed a set of students' essays and picked up the following five frequent serious errors that should be detected:

(a) Spelling errors

(b) Inappropriate choice of case particles

(c) Nouns with inappropriate modifiers

(d) Inappropriate letter choice in nouns
(mischoices among hiragana, katakana, and kanji)

(e) ordering errors of phrases (*bunsetsu* in Japanese)

Then we prepared an artificial set of data in which five types of errors pointed above were mechanically added. With this data set, we tested two error detection models for Japanese optical character reading ([5], [6]), and

made experiments for evaluating these two models and tuning parameters. One model ([5]) focuses on the contextual allegation and detects the maximum inappropriateness. The other model ([6]) employs *m*th Order Markov Model to detect the lowest probability of transitions of a given string x_i following the prior *m* strings using the following equation:

$$P(x_i \mid x_{i-m}, x_{i-m+1}, \cdots, x_{i-1}) \equiv \frac{O(x_{i-m}, x_{i-m+1}, \cdots, x_{i-1}, x_i)}{O(x_{i-m}, x_{i-m+1}, \cdots, x_{i-1})}$$

Neither models can detect all the errors of (a) to (e) but tuning up the parameters with the artificial set of data improved the detection rate.

As for the second problem, it is desirable to visualize only the appearances of undesirable expressions effectively, but at the same time, visualizing the sequence of the raw value of gr (*e*) as in Figure 3 should be smoothed. For these purpose, we propose a function score (x, w) for smooth visualization of the appearances of undesirable expressions for the targeted subgenre:

$$\text{score (x, w)} = -\sum_{e \in \text{n-gram (x,x+w)}} \text{b(e)}$$

where b (e) = 0 when dis (e) – α >0
= (dis (e) – α)^2 when dis (e) – α <0
α is a heuristically determined coefficient.
n-gram (x, x+w) is a function that returns the set of n-gram expressions included from x-th to w-th strings.

With this function, the result visualization of the same student's essay used for Figure 3 is as follows:

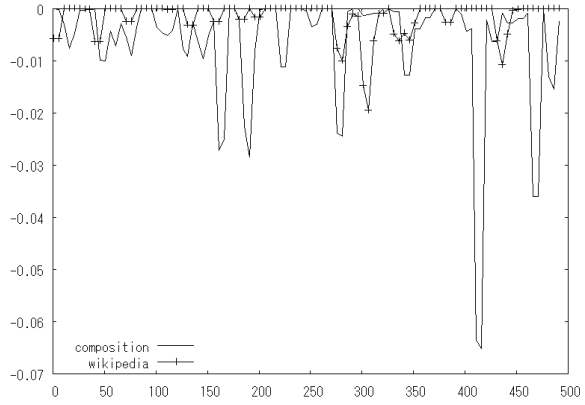

Fig. 4 Smoothed Visualization

With these improvements, we are currently implementing a pilot application for enabling teachers and students to easily detect expressions that are to be revised or improved in an essay.

V. CONCLUSION

In this paper, as a foundational data accumulation and analyses for them for the purpose of developing educational tools for improving stylistic aspects of writing skills, we proposed a method using multi-corpora comparison to correctly extract expressions not suited to a particular subgenre intended, and developed an experimental visualization for teachers and students. Much has to be done towards a practically effective tool, but any advanced tool to help teachers and students to detect undesirable expression should be conscious of stylistic differences among subgenres and our proposed methods are foundationally effective.

REFERENCES

[1] Strunk Jr., W. and E.B. White (2000[4]) *The Elements of Style*, Longman.
[2] Williams, J.M. (2008) *Style: The Basics of Clarity and Grace*, Longman.
[3] Japanese Wikipedia dumped file http://download.wikimedia.org/jawiki/20071013/jawiki-20071013-pages-articles.xml.bz2
[4] Masao Uchiyama and Kiyomi Chujo (2007) Linking Word Distribution to Technical Vocabulary, Technical Report B, College of Industrial Technology, Nihon Univ. Vol.40 pp.13-21.
[5] Masaki Murata, Hitoshi Isahara (2002) Automatic Detection of Mis-Spelled Japanese Expressions Using a New Method for Automatic Extraction of Negative Examples Based on Positive Examples, IEICE Transactions on Information and Systems, E85-D(9) pp.14165-1424.
[6] Tetsuo Araki, et al, (2000) A Method for Detecting and Correcting of Characters Wrongly Substituted, Deleted, or Inserted in Japanese Strings Using *m*th-Order Markov Model, The Transactions of the Institute of Electronics, Information and Communication Engineers, D-II J83-D-II(6) pp.1516-1528.