Semiautomatic Interactive Multilingual Style Analysis (SIMSA)

Günter Winkelmann
TA Triumph-Adler AG
TA Research
Fürther Strasse 212
D-8500 Nuremberg 80

1 Abstract

A style checker is a tool which supports authors during the process of writing: Certain style markers are analyzed, their values are compared with a given norm, deviations are detected, and recommendations are given to the author. The power of a style checker depends on available tools such as lexica, parser, etc. This paper describes a style checker which will be integrated in a workbench (Translator's Workbench) and which has access to lexica and parser. The style checker can be used for different languages and for different kinds of text.

2 Introduction

Within ESPRIT II - Project 2315, Translator's Workbench (TWB), a tool is under development which checks stylistic markers of texts. According to the goals of Translator's Workbench as an integrated multilingual toolkit, the concept of the style checker includes multilinguality, and it uses other tools of Translator's Workbench such as parser and lexicon, which provide SIMSA with more power than comparable approaches.

Style as it is understood in SIMSA is not reduced to simply a personal impression. Its description is not restricted to informal classifications such as "good style", "style like the style of author x". Within a functional, group based definition, style is the selection of certain words, phrases, sentences or structures out of a set of grammatically correct words, phrases, sentences or structures. These linguistic features in which texts can differ stylistically are called style markers. Group based means that these style markers have equal characteristics within a certain group of texts.

The stylistic characteristics of a group of texts, their values of style markers, can be set as a norm. Thus, different norms can be defined with the help of these values: A stylistic norm for technical texts, for user manuals, for a certain author, etc.

With respect to these premises we now define style checking as: matching the densities of the features in a given text against the densities of the corresponding features in the norm. "Correct style" is correct style concerning a certain (group based) norm. "Incorrect style" is the degree of deviation between the norm values of the style markers and the values in the actual text.

3 Recent Approaches

Similar approaches (but within different contexts) have been done during the last decade. Beside approaches translating style from one language to another via abstract universal categories like text complexity and readability (Dimarco/Hirst 1988), two approaches to style checking, Writer's Workbench and EPISTLE, should be mentioned.

Writer's Workbench (Cherry 1983 et al.) has influenced the development of commercial software such as Rightwriter, PC-Style and Grammatik. It does style critiquing, but it cannot do critiquing that requires a parser output, such as noun phrase complexity. Embedded in EPISTLE (Heidorn et al. 1982) is a style checker which uses the parse tree of the grammar check. So, EPISTLE covers a wider range of style markers. Nevertheless, stylistic critiques have to be adapted to the field of application.

Compared with Writer's Workbench and EPISTLE, SIMSA is a more universal approach. Its main feature is multilinguality. The same parser can be used for different languages (cf. Hellwig 1988), which allows to use the same format of parser output for the style analysis of different languages. Additionally, style markers are more universal and can be used in different languages too, as well as in different kinds of text. The adaptation to a language or a special kind of text will be done automatically if a sufficiently large text corpus is put in for the setting of a new norm ("Sufficiently large" means sufficient for significant values concerning each feature; therefore "sufficiently large" depends on the selected features).

4 Style markers

On the very beginning of style analysis, we need an inventory of style markers.

Style errors can be detected on several different levels: word, phrase, sentence and text.

Relevant stylistic features are

- on word level: word length; fillers; nominalisation; compound nouns, terminology;
- on phrase level: noun-phrase complexity; cumulation of adjectives; complex prepositional phrases;
- on sentence level: sentence length; compound sentences; distance between verb stem and prefix;
- on text level: passive voice; pronouns; phenomena of cohesion/ coherence: reference, conjunctions, etc.

Within the project, two teams (Siemens AG, Germany and Triumph-Adler AG, Germany) are working on the development of relevant style markers. The development is conducted in four steps.

First, principles of good style and possible stylistic markers in general had to be identified by examining literature on good technical writing and linguistic literature on style markers.

For each style marker the information needed has been identified, so that it can be used by the style checker. Some style markers can be transferred into an algorithm just by using statistical methods, others need lexical information, and a third group needs syntactic information which has to be provided by the parser within the TWB project.

In a third step, the style markers are formalized and checking algorithms are being developed.

Finally, functions are being developed to transfer given values (average, standard deviation, etc) in a bar chart representation including thresholds and the degree of deviation from the predefined norm.

5 Architecture

SIMSA consists of three main parts. The user has the option to set the norms and thresholds of the stylistic features by putting in a representative or paradigmatic text corpus (standardization of style marker values). He can perform an analysis of a given text (Analysis; batch mode) and he can start a dialogue for more information on a given analysis (Analysis dialogue; interactive mode).

Standardization of style marker values

Importance of style markers, their average values, and thresholds of their values depend on the analyzed language, and they differ with the kind of analyzed texts. How can stylistic critiques be adapted to different fields of application? In principle, there are three possibilities:

First, stilistic norms can be fixed once and for all without any possibility of change. This case allows only one conception of "one good style". But what about functional concepts in which deviations concerning style markers are understood as deviations in the functionality of a text? And what about different functionalities of style markers in different languages?

A second approach is to set the standard norm by the users themselves or at least by a superuser. This is the approach in EPISTLE where "thresholding, together with adjustable weights, allow tailoring of style critiques to individual environments..." (Heidorn et al 1982:323).

A somewhat different approach was taken in SIMSA: SIMSA provides the user with default norms for several kinds of text. Moreover, it offers the option to set norms according to a given text corpus. The user puts in some texts which belong to a given language and a given kind of text. SIMSA will analyze the text corpus and will set and store the norm of the style markers accordingly.

Analysis

The analysis part of SIMSA is under development by the above mentioned teams of Siemens AG and TA Triumph-Adler. Due to the nature of TWB as an integrated toolkit, analysis functions will use other TWB tools as far as possible. The analysis functions can be divided into three main groups, in purely statistical functions, in functions with lexical access, and in functions using parser output.

Statistical algorithms are sufficient for style markers as e.g. sentence length and word length. The analysis functions check the size of the text corpus (a certain size is necessary to get significant deviations), compute average, standard deviation and other necessary values and compare these values with the norm values.

Functions using lexical information are necessary for style markers as e.g. fillers and slang expressions. The access to lexica can be managed in two ways. Either words can be matched against small lexica specially designed for stylistic purposes containing only a small amount and semantically restricted class of words (e.g. fillers or slang expressions), or words can be matched via the parser output against the lexicon used by the parser. In the second case, necessary stylistic information (e.g. "word is a chemical technical term") is contained in the lexicon entry.

Functions using parser output are necessary for style markers as e.g. noun phrase complexity, distance between verb stem and verb prefix, sentence complexity. These functions filter the parser output for necessary information.

The style checker still works if parser access is not possible. In this case (and in cases the user doesn't want an analysis concerning all style markers) the analysis of certain style markers can be suppressed.

The results of the analysis (values of deviation from the norm, etc.) are stored in a separate analysis file.

Analysis dialogue

There are two ways to start the analysis dialogue. First, an option "analysis dialogue" will be offered to the user after the style checker has finished its analysis. Second, the user can call the analysis dialogue separately if there is an analysis file and a corresponding text file.

"Analysis dialogue" opens a window containing bar charts which demonstrate for each analyzed style marker the degree of deviation from the norm. The user can ask for more information about certain style features in general and he can ask for the occurrences of the criticized style markers in the text. Due to the nature of stylistic errors as grammatically correct but more or less inadequate usage of linguistic features, the "Analysis dialogue" is thought to give recommendation as far as possible, but not to correct text passages automatically.

6 References

L.L. Cherry et al.: Computer aids for text analysis. in: Bell Laboratories Record. Volume 61, Number 5, Short Hills, New Jersey, 1983, 10-16.

C. Dimarco/G. Hirst: Stylistic Grammars in Language Translation. in: Proceedings of the 12th International Conference on Computational Linguistics (COLING), Budapest, 1988, 148-153.

G.E. Heidorn et al.: The EPISTLE text-critiquing system. in: IBM System Journal, Vol. 21, Nr. 3. 1982, 305-326.

P. Hellwig: Chart parsing according to the slot and filler principle. in: Proceedings of the 12th International Conference on Computational Linguistics (COLING), Budapest, 1988, 242-244.

G. Heyer, R. Kese, M. Lüdtke and G. Winkelmann: Translator's Workbench - A toolkit for translators. in: ESPRIT '89. Office and Business Systems. Results and Progress of Esprit Projects in 1989. Brussels, November 1989.

Enkvist, N. E.: Linguistic Stylistics. Paris, 1973.