Automatic Identification of Organisational Structure in Writing using Machine Learning

Laurence Anthony and George V. Lashkia

Dept. of Information and Computer Engineering

Okayama University of Science

1-1 Ridai-cho, Okayama 700-0005, Japan

anthony@ice.ous.ac.jp lashkia@ice.ous.ac.jp

ABSTRACT (150-200 words) 175 words

Teaching learners about the common structural patterns used in different types of texts, such as the abstract and introduction of research papers, has proved successful in many ESP reading and writing courses. However, a major problem faced by researchers when analyzing texts is the vast amount of time needed to conduct the analysis. This has led to many studies reporting only 'preliminary' findings, based on a small corpus of target texts.

In this paper, we propose a computer system that uses machine learning to automatically identify the structure of texts, enabling researchers to quickly and effectively process very large corpora. The system also has applications in the classroom as a teacher resource when evaluating and selecting texts that highlight certain features, and as a student resource when conducting data-driven learning.

To test the system, it was applied to abstracts from computer science journals and found to be fast and accurate. It was also assessed by practicing ESP teachers and learners and shown to be flexible, easy to use, and a practical aid in the classroom.

Automatic Identification of Text Structure in Writing using Machine Learning

Laurence Anthony and George V. Lashkia

Dept. of Information and Computer Engineering

Okayama University of Science

1-1 Ridai-cho, Okayama 700-0005, Japan

anthony@ice.ous.ac.jp lashkia@ice.ous.ac.jp

DETAILED SUMMARY (500 words) 484 words

Research has shown that teaching the common structural patterns used in different types of texts can be particularly useful for second language (SL) and foreign language (FL) learners, who may have writing difficulties at the sentence level compounded by a lack of knowledge or experience at the discourse level. The interest in text structure has led to a large number of studies endeavoring to discover the characteristic structures of writing in different fields and text types, such as the different sections of the research article, and the structure of patents, grant proposals, law reports, and so on.

Much of the research on text structure has been based on a quantitative analysis of a corpus of authentic target texts. However, for a corpus to accurately represent the target language it has to be fairly large, and the analysis of structure in a large-scale corpora is inherently a time consuming process. Unfortunately, this has led to many researchers concentrating on only a small corpus of target texts, and reporting only 'preliminary' findings. The rapid advances in computer technology have also offered little to the structural analyst, with most of the current systems focusing instead on more 'bottom-up' sentence level tasks, such as searching, counting, and sorting of linguistic items. Subsequently, there is a clear need for more advanced programs that can assist researchers at this 'top-down' discourse level.

In this paper, we propose such a system that uses machine learning to automatically identify the structure of texts. To achieve this, the system first 'learns' the characteristic features of structural steps in a particular discipline or field based on a statistical analysis of a small number of hand classified examples. When given a target text, the system divides this into sentences which are then analyzed to determine which of the features it has learnt are present or not. Finally, based on these results the most probable structural step for each sentence is decided, and the target text labeled

accordingly.

To evaluate the performance of the system, it was given the task of analyzing the structure of abstracts from computer science journals and the labeling it produced compared with that of a trained analyst. Results showed that the system could generate an average first order accuracy of approximately 70%, but this could be improved to almost 90% by including second order probability decisions. The system was also tested by practicing ESP teachers and learners who were given typical classroom tasks to solve with and without the aid of the system. From these experiments, it was shown that the system could be a practical aid to the teacher, by greatly reducing the time required to evaluate and select texts for use as classroom materials. It also served as useful student resource when conducting data-driven learning on the structure of texts. Finally, the system was shown to be flexible and easy to use, incorporating a full windows-based graphical user interface.

BIOGRAPHY (20-30 words)

Laurence Anthony has a B.S. degree in mathematical physics and a M.A. degree in TESL/TEFL. He is currently completing his Ph.D. in applied linguists at Birmingham University, UK.