

Semantic Coherence Analysis of English Texts Based on Sentence Semantic Graphs

Nanxiao Deng¹, Yabing Wang¹, Guimin Huang^{1,2*}, Ya Zhou¹ and Yiqun Li¹

¹School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, Guangxi, China

²Guangxi Key Laboratory of Image and Graphic Intelligent Processing

Abstract

With the reform of China's education industry, more and more universities are using computers to conduct examinations. For the automatic correction of essays as subjective questions, existing automatic English text scoring systems suffer from insufficient extraction of coherence information and low accuracy when analysing text coherence. Therefore, this paper proposes an unsupervised semantic coherence analysis model for English texts based on sentence semantic graphs, taking Chinese students' English compositions as the research context. Guided by the semantic coherence theory, the English text is represented as a sentence semantic graph, and an improved VF2 subgraph matching algorithm is used to mine the frequently occurring subgraph patterns in the sentence semantic graph. After that, the set of frequent subgraphs is generated by filtering the subgraph patterns according to their frequencies, and the subgraph frequency of each frequent subgraph is calculated separately. Finally, the distribution characteristics of frequent subgraphs and the semantic values of subgraphs in the sentence semantic graphs are extracted to quantify the overall coherence quality of English texts. The experimental results show that the model proposed in this paper has higher accuracy and practical value compared with the current methods of coherence analysis.

Keywords: english text, semantic coherence theory, sentence semantic graph, VF2 subgraph matching algorithm, frequent subgraph

Received on 03 May 2023, accepted on 27 August 2023, published on 28 August 2023

Copyright © 2023 N. Deng *et al.*, licensed to EAI. This is an open access article distributed under the terms of the [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/), which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi: 10.4108/eetsis.3312

*Corresponding author. Email: sendhuang@126.com

1. Introduction

In recent years, techniques related to natural language processing have become more and more widely used. As an important application in the field of natural language processing, the detection and evaluation of semantic coherence has developed rapidly [1-3]. There is an urgent need for researchers to assess the coherence quality of the large number of textual results generated by many intelligent systems, such as the results of automatic abstract generation and machine translation, because if the coherence of these texts is poor, it will create a significant barrier to reading and even lead to incomprehension of the meaning of the text. In addition, the quality of coherence is an important

criterion in all English composition scoring systems, and its analysis is essential. As a result, researchers have begun to investigate and quantify the quality of text coherence with a view to its practical application, and this has led to the study of text coherence. Generally speaking, when scoring English language learners' essays, the scoring criteria should cover four aspects: lexical complexity, grammatical accuracy, syntactic complexity and discourse coherence [4], which are the prerequisites for accurate and reliable scoring results. However, existing automatic English essay scoring systems rarely address the indicator of coherence, which results in an unreasonable final score for English essays. For example, when a large number of sentences are inserted into an essay with excellent lexical complexity, grammatical accuracy and syntactic complexity, the final score will be high even if the overall coherence of the essay is not high. This increases the

