

## Preface

In recent years, as Chinese has become an increasingly popular second language, researchers have focused more on assisting Chinese language learning. One significant area of this research is the automatic Chinese grammatical error correction on sentence-level. However, most studies emphasize constructing neural network models while neglecting the unique features and error patterns inherent to the Chinese language. This book addresses this gap by analyzing Chinese interlanguage corpora to examine automatic correction at the character, word, and sentence levels. The primary research content includes:

First, this book conducted a detailed statistical analysis of errors in Chinese sentences using the HSK Dynamic Composition Corpus. The analysis revealed that errors at the character and word levels occurred frequently, accounting for 87.65% of the total errors. Many of these errors were concentrated on a specific set of characters and words, indicating high overlap and recurrence rates. As a result, this book compiled two dictionaries from the corpus: a context-independent dictionary and a context-dependent dictionary. The context-independent dictionary allows for the direct correction of erroneous words without considering the context, while the context-dependent dictionary provides a high-frequency candidate list for each target word based on contextual information.

Second, this book explored a multi-layer representation method that integrates grammatic and word order information on character and word-level. Based on context information, word order information, and random permutation, this method learned the context and word order information of characters and words from large-scale corpora, incorporating them into word embedding. To vali-

date the effectiveness of the word embedding, this book trained character and word embedding on a small-scale corpus and evaluated their performance through a word ordering task to assess the ability of word embedding to capture the word order information in sentences.

Then, this book constructs a test corpus for predicting Chinese sentence acceptability (CGUG) based on an English acceptability evaluation corpus (GUG), using interlanguage corpora. The research on sentence acceptability can help in scoring generated sentences, making them closer to native speaker standards. The book finds that a high-order N-gram language model, combined with effective smoothing algorithms, spelling error handling, and word frequency, can accurately predict sentence acceptability. Additionally, it explores the relationship between the language model perplexity and sentence acceptability, revealing that even language models with higher perplexity can still effectively predict sentence acceptability.

At last, this book develops a grammatical error correction model on sentence-level integrating Chinese error patterns and sentence acceptability. Initially, an improved Transformer model is used to correct target sentences. To enhance accuracy and recall, the model incorporates error-free sentences and classification information, enabling it to better learn Chinese sentence patterns. Subsequently, character-level and word-level errors are corrected using context-independent and context-dependent dictionaries. Finally, the model ranks the output candidate sentences based on sentence acceptability. Experimental results demonstrate that this integrated model effectively corrects Chinese sentence errors by combining error patterns and sentence acceptability.

The contributions and innovations of this book are primarily

demonstrated in the following aspects: this book proposes an improved sentence acceptability algorithm for error analysis. It refines existing methods by examining sentence acceptability from the perspectives of word compatibility, word order, and language models. The findings indicate that a high-order N-gram language model, combined with effective smoothing algorithms, character-level errors, and word frequency, can accurately predict sentence acceptability. On the existing GUG test set, the model achieved a correlation of  $r = 0.522$ , which is 5 percentage points higher than the previous best result of 0.472.

This book introduces an automatic grammatical error correction model on sentence-level based on deep learning. For the first time in error correction tasks, native Chinese corpus (literary works) and sentence classification information (acceptable or unacceptable) were incorporated into the Transformer model for training. Previous research often overlooked the significant impact of native language information on error correction. Experiments demonstrated a substantial improvement in model accuracy, increasing from 9.7% to 48% with the inclusion of native corpus and classification information.

This book proposed a comprehensive method that integrates deep learning algorithms, Chinese character and word error patterns, and a sentence acceptability model. Building on deep learning algorithms, this method uses Chinese character and word error patterns for further correction and employs the sentence acceptability model for ranking the results. The results indicate that integrating Chinese character and word error patterns can enhance the accuracy and recall of deep learning models, increasing the F1 score from 29.97% to 31.86%. Additionally, the sentence acceptability model improved the binary classification accuracy (acceptable or unacceptable) from 87.9% to 90.1%.