Contribution ID: **84** Type: **not specified**

Construction of an Evaluation Model for Language Corpora

The growing demand for Vietnamese natural language processing applications highlights the urgent need for high-quality corpora; however, no unified reference framework currently exists for evaluating such resources. This study proposes a model for evaluating Vietnamese corpora by combining quantitative metrics (lexical richness, frequency) and qualitative aspects (coverage, sentence structure, word usage accuracy), supported by WordSketch for collocation analysis. The proposed model demonstrates the feasibility of integrating quantitative and qualitative criteria into a unified process that can systematically reflect both the quality and limitations of the corpus, while also showing potential applicability to Vietnamese. This framework contributes to establishing a reference for corpus evaluation, opening pathways for improving and expanding Vietnamese linguistic resources to support future NLP research and applications.

Từ khóa

quality corpora, corpus evaluation, corpus quality evaluation, collocation, evaluation

Thông tin các tác giả

1. Vũ Thi Thi: CN., đang công tác tại Trường Đại học Khoa học tự nhiên, ĐHQG-HCM, TK16/19 Nguyễn Cảnh Chân, phường Cầu Ông Lãnh, TP. HCM, email: vtthi@hcmus.edu.vn

Author: VU, Thi-Thi

Track Classification: Tiểu ban 1: Những tiến bộ và thành tựu mới trong lĩnh vực Ngôn ngữ học Tính toán