GIẢI PHÁP XÂY DỰNG KHO NGỮ LIỆU ĐA NGỮ VIỆT-ÊĐÊ GÁN NHÃN THEO NGỮ CẢNH

SOLUTIONS TO BUILDING THE VIET - EDE MUTILINGUAL CORPUS WITH THE CONTEXTUAL LABEL

Tác giả: *Hoàng Thi Mỹ Lê*, *Phan Huy Khánh*

Tóm tắt bằng tiếng Việt:

Trong lĩnh vực xử lý ngôn ngữ tự nhiên (XLNNTN), kho ngữ liệu đa ngữ là một tài nguyên rất cần thiết. Chất lượng của kho ngữ liệu đa ngữ đóng vai trò quyết định đến chất lượng đầu ra của hệ dịch. Hệ dịch sẽ không cho kết quả tốt nếu kho ngữ liệu đa ngữ sử dụng trong quá trình huấn luyện có chất lượng không tốt cho dù được áp dụng các phương pháp học máy tiên tiến nhất. Hiện nay chưa có một kho ngữ liệu song ngữ Việt-ÊĐê với phông chữ Unicode nào đã được công bố chính thức và cho phép cộng đồng nghiên cứu có thể chia sẽ sử dụng để nghiên cứu. Từ đó, bài báo đề xuất giải pháp xây dựng kho ngữ liệu đa ngữ Việt-ÊĐê với phông chữ Unicode có xử lý nhập nhằng và từ đa ngữ nghĩa, bằng cách gán nhãn theo từng ngữ cảnh thuộc lĩnh vực giáo dục như giáo dục về chăn nuôi, trồng trọt, bảo vệ rừng, chăm sóc sức khoẻ, v.v… cho các đồng bào các dân tộc thiểu số Việt Nam.

Từ khóa: Kho ngữ liệu đa ngữ; Dân tộc thiểu số; ÊĐê; Unicode; Tách từ

Tóm tắt bằng tiếng Anh:

In the natural language processing (NLP), the multilingual corpus is a necessary resource. The quality of multilingual corpus plays a decisive role in the output quality of the translational system. The translational system will not produce a good output, if the the quality of multilingual corpus in the training process is not good, though the most advanced machine learning methods are applied. Currently, there is no Vietnamese-EDe multilingual corpus using Unicode fonts, which has been officially announced and allows the research community to share and use for research purposes. For this reason, the propose of this paper is to develop a solution to building a Vietnamese-EDe multilingual corpus using the Unicode font which can process the ambiguity and multi - meaning words by labeling each word with the context in the educational field such as education in animal husbandry, cultivation, forest preservation, health care, etc.... for the ethnic minorities (EM) in Vietnam.

Key words: multilingual corpus; the ethnic minorities; Ede; Unicode; word segmentation