GIẢI PHÁP XÂY DỰNG NHANH NGỮ LIỆU SONG NGỮ TỪ TED

***Hoàng Khuê*[1], *Lương An Vinh*[2]**

[1]Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM

[2]Khoa Công nghệ Thông tin, Đại học Công nghệ Sài Gòn

[1]hkhueedu@gmail.com, [2]vinh.luongan@stu.edu.vn

Tóm tắt:

Hiện nay, ngữ liệu song ngữ đóng vai trò quan trọng đối với việc nghiên cứu và xử lý ngôn ngữ tự nhiên. Tuy nhiên, việc tự xây dựng hoặc mua ngữ liệu song ngữ có sẵn với chất lượng tốt sẽ tốn nhiều chi phí, đặc biệt đối với các cặp ngôn ngữ nghèo tài nguyên như tiếng Việt (vi) – Quốc Tế Ngữ (eo). Vì vậy, trong bài viết này, chúng tôi sẽ trình bày ý tưởng sử dụng tiếng Anh, một ngôn ngữ giàu tài nguyên, làm trung gian để xây dựng ngữ liệu song ngữ cho bất kì cặp ngôn ngữ nào từ dữ liệu của TED[i]:

$$\text{TED} \xrightarrow{tải\ dữ\ liệu\ en-L2} \{\text{en-vi;en-eo}\} \xrightarrow{\{vi-en\}\ so\ khớp\ \{en-eo\} với\ trung\ gian\ en} \{\text{vi-eo}\}$$

Bằng giải pháp trên, chúng tôi đã xây dựng được một kho ngữ liệu song ngữ gồm 7 cặp ngôn ngữ: vi-fr; vi-zh_cn; vi-ja; vi-ko; vi-de; vi-ru; vi-eo, với tổng số 791056 cặp câu.

Từ khóa: ngữ liệu, song ngữ, TED

[i] TED là tổ chức phi lợi nhuận hoạt động với mục tiêu truyền bá các ý tưởng thông qua các buổi nói chuyện ngắn (18 phút đổ lại). TED bắt đầu từ năm 1984 như một hội nghị quy tụ các chủ đề về công nghệ giải trí và thiết kế. Ngày nay thì các chủ đề mỡ rộng ra mọi lĩnh vực từ khoa học đến kinh doanh bằng hơn 100 ngôn ngữ. https://www.ted.com/about/our-organization

BUILDING QUICKLY PARALLEL CORPUS FROM TED

***Hoàng Khuê*[1], *Lương An Vinh*[2]**

[1]University of Science, VNU-HCM

[2] Saigon Technology University, STU

[1]hkhueedu@gmail.com, [2]vinh.luongan@stu.edu.vn

Abstract:

Nowadays, parallel corpora play an important role in researching and processing natural language. However, building or buying an available qualified parallel corpus may be costly, especially low resource language pairs like Vietnamese (vi) – Esperanto (eo). There fore, in this article, we will present the idea of using English, a rich resource language, as an intermediate language for building parallel corpus for any language pairs from TED[i] data.

$$\text{TED} \xrightarrow{download\ \{en-L2\}} \{\text{en-vi;en-eo}\} \xrightarrow[with\ en\ is\ \text{intermediate}]{\{vi-en\}\ matched\ \{en-eo\}} \{\text{vi-eo}\}$$

With this solution, we built a parallel corpus consisting of seven language pairs: vi-fr; vi-zh_cn; vi-ja; vi-ko; vi-de; vi-ru; vi-eo, with 791.056 pair sentences in total.

Key word: corpus, parallel, TED

[i] TED is a nonprofit devoted to spreading ideas, usually in the form of short, powerful talks (18 minutes or less). TED began in 1984 as a conference where Technology, Entertainment and Design converged, and today covers almost all topics — from science to business to global issues — in more than 100 languages. https://www.ted.com/about/our-organization