DOI: 10.15625/vap.2016.000105

XÂY DỰNG TỪ ĐIỂN MỚI BẰNG PHƯƠNG PHÁP ÁNH XẠ TRUNG GIAN

Khang Nhứt Lâm

Khoa Công nghệ thông tin và Truyền thông
Trường Đại học Cần Thơ
lnkhang@cit.ctu.edu.vn

TÓM TẮT— *Để xây dựng một từ điển mới từ một ngôn ngữ nguồn A sang một ngôn ngữ đích C, ký hiệu là Dict(A,C), các phương pháp hiện tại cần nhiều nguồn tài nguyên có sẵn như các tự điển trung gian hoặc một lượng lớn tài liệu văn bản ở ngôn ngữ A, C và ở một hoặc nhiều ngôn ngữ trung gian B. Tuy nhiên, không phải ngôn ngữ nào cũng có các nguồn tài nguyên sẵn có như vậy, đặc biệt là các ngôn ngữ ít tài nguyên hoặc ngôn ngữ "nguy cấp" được UNESCO thống kê. Bài báo này trình bày phương pháp làm tăng số lượng entry trong từ điển sẵn có và xây dựng một từ điển mới Dict(A,C) bằng phương pháp ánh xạ trung gian từ 2 tự điển sẵn có Dict(A,B) và Dict(B,C) với B là một ngôn ngữ phổ biến. Cụ thể hơn, chúng tôi sử dụng tiếng Anh như một ngôn ngữ trung gian để thực hiện việc ánh xạ ngữ nghĩa các từ vựng ở ngôn ngữ A sang ngôn ngữ C. Để loại bỏ các mục dịch không chính xác, chúng tôi sử dụng phương pháp tính trọng số. Nếu một mục dịch có trọng lượng lớn hơn một ngưỡng δ, nó được xem như là mục dịch tốt và sẽ được thêm vào tự điển mới Dict(A,C). Nghiên cứu của chúng tôi không chỉ giúp giảm đáng kể chi phí xây dựng các tự điển mới, mà nó còn góp phần hỗ trợ các cộng đồng sử dụng ngôn ngữ không có nhiều nguồn tài nguyên từ vựng.*

Từ khóa— *Tự điển, ánh xạ trung gian, mục dịch, ngôn ngữ nguy cấp*

I. GIỚI THIỆU

Tổ chức Ethnologue[1] thống kê có hơn 7,106 ngôn ngữ đang tồn tại trên thế giới. Phần lớn các ngôn ngữ không có nhiều nguồn tài nguyên từ vựng như từ điển, thesauri, cơ sở dữ liệu từ vựng (Wordnet) và tập các văn bản (corpora). Các từ điển chúng ta đang có đa phần là giữa các ngôn ngữ phổ biến (ví dụ: từ điển giữa các ngôn ngữ như Anh, Tây Ban Nha và Đức) hoặc giữa một ngôn ngữ phổ biến và một ngôn ngữ ít phổ biến hơn (ví dụ: từ điển Anh-Việt, Pháp-Ả Rập và Đức-Lào). Từ điển giữa các ngôn ngữ có ít nguồn tài nguyên (resource poor languages) hoặc các ngôn ngữ "nguy cấp" (endangered languages) chiếm số lượng rất hạn chế, thậm chí là không có. Ví dụ, chúng ta có thể tìm thấy từ điển Assamese - Anh và tự điển Anh - Việt, nhưng tự điển Assamese - Việt là chưa tồn tại. Assamese là một ngôn ngữ Indo-European ở Ấn Độ, được sử dụng bởi khoảng 30 triệu người, nhưng là một ngôn ngữ có rất ít tài nguyên từ vựng. Rất nhiều ngôn ngữ gần như phải rất may mắn chúng ta mới tìm được một từ điển giữa nó và một ngôn ngữ phổ biến. Trường Đại học Chicago[2] cung cấp các tự điển song ngữ cho 29 ngôn ngữ ở các quốc gia Đông Nam Châu Á, nhưng nhiều ngôn ngữ trong số này chỉ có duy nhất một từ điển. Các từ điển hiện có cũng rất khác nhau về chất lượng và số lượng từ trong từ điển. Từ điển song ngữ không chỉ là một tài nguyên từ vựng của một ngôn ngữ nào đó, mà nó còn là yếu tố sống còn của một ngôn ngữ. Để một ngôn ngữ được tồn tại và phát triển mãnh mẽ thì ngôn ngữ đó phải được sử dụng không chỉ trong giao tiếp hàng ngày mà còn sử dụng trong các khía cạnh khác của cuộc sống như học thuật, nghiên cứu và kinh doanh. Để xây dựng một từ điển có chất lượng cao thì chúng ta cần nhiều tài nguyên khác hỗ trợ, chẳng hạn như các từ điển trung gian, Wordnet hoặc corpora. Mặt khác, để xây dựng được các cơ sở dữ liệu Wordnet và corpora có chất lượng cần phải có các từ điển. Có thể nói từ điển song ngữ là một trong những tài nguyên từ vựng rất cần thiết để xây dựng các tài nguyên từ vựng khác. Xuất phát từ nhu cầu thực tiễn, chúng tôi tìm hiểu phương pháp để xây dựng từ điển song ngữ mới.

Mục tiêu nghiên cứu của chúng tôi là từ các từ điển sẵn có của các ngôn ngữ (i) chúng tôi sẽ làm tăng số lượng entry có trong từ điển sẵn có, (ii) xây dựng các từ điển song ngữ mới cho các ngôn ngữ đó với độ chính xác không quá thấp hơn độ chính xác từ điển sẵn có, (iii) phương pháp chúng tôi giới thiệu phải có khả năng áp dụng được cho các ngôn ngữ không có nhiều nguồn tài nguyên. Cụ thể, mục II sẽ trình bày cấu trúc từ điển. Các tài liệu liên quan được đề cập trong mục III. Mục IV giới thiệu về các ngôn ngữ và từ điển song ngữ sẽ được sử dụng. Phương pháp làm tăng số lượng entry trong từ điển và xây dựng từ điển mới được trình bày trong mục V. Chúng tôi sẽ trình bày kết quả thực nghiệm và thảo luận trong mục VI. Cuối cùng mục VII sẽ tổng kết nghiên cứu của chúng tôi.

II. CẤU TRÚC TỪ ĐIỂN

Trước khi giới thiệu phương pháp xây dựng từ điển mới từ các từ điển song ngữ sẵn có, chúng tôi sẽ giới thiệu về cấu trúc của một từ điển. Một từ điển song ngữ A-B chứa các mục dịch hay còn gọi là các "entry" dịch các từ hoặc cụm từ ở ngôn ngữ nguồn A sang các từ hoặc cụm từ ở ngôn ngữ đích B. Một từ điển song ngữ A-B, ký hiệu là *Dict(A,B)*, khác với một tự điển song ngữ B-A, ký hiệu là *Dict(B,A)*. Cụ thể hơn, *Dict(A,B)* chứa các entry *(a,b)*, trong khi *Dict(B,A)* chứa các entry *(b,a)*. Một entry trong từ điển, còn được gọi là *LexicalEntry* có dạng <*LexicalUnit, Definition*>. Theo Landau [1], một *LexicalUnit* là một từ hoặc một cụm từ sẽ được định nghĩa. Nói cách khác, một từ điển là một danh sách các *LexicalEntry* được sắp xếp theo thứ tự dựa trên các *LexicalUnit*. Với một *LexicalUnit*, phần *Definition* tương ứng của nó thường bao gồm loại từ (Part-Of-Speech - *POS*), cách phát âm, nghĩa (*sense*), ví dụ minh

[1] https://www.ethnologue.com/
[2] http://dsal.uchicago.edu/dictionaries/list.html

họa sử dụng từ trong ngôn ngữ nguồn và ngôn ngữ đích, và một số thông tin khác. Một *LexicalUnit* có thể có nhiều hơn một *sense*. Do đó, một entry trong tự điển có dạng <*LexicalUnit, Sense*$_1$*, Sense*$_2$*,…*>.

III. TÀI LIỆU LIÊN QUAN

Giả sử tồn tại tự điển *Dict(A,B)* chứa các entry $(a_i;b_k)$ và tự điển *Dict(B,C)* chứa các entry (b_k,c_j). Các từ trong mỗi entry ở cả ngôn ngữ nguồn a_i, ngôn ngữ trung gian b_k và ngôn ngữ đích c_j có thể là một từ đơn, từ ghép hoặc cụm từ. Phương pháp "ngây thơ" (naïve approach) xây dựng một tự điển mới *Dict(A,C)* tiến hành như sau: nếu từ a_i ở ngôn ngữ A có nghĩa là từ b_k ở ngôn ngữ B và từ b_k có nghĩa là c_j ở ngôn ngữ C, thì phương pháp "ngây thơ" đưa ra kết luận là từ a_i ở ngôn ngữ A có nghĩa là c_j ở ngôn ngữ C. Tuy nhiên, nếu b_k có nhiều hơn một nghĩa thì phương pháp này sẽ đưa ra những kết luận sai, đây được gọi là sự nhập nhằng ngữ nghĩa (Word Sense Disambiguation - WSD). Nhiều phương pháp được giới thiệu để loại bỏ vấn đề nhập nhằng ngữ nghĩa như sử dụng thông tin từ các tự điển trung gian khác sẵn có [2], [3], [4] hoặc thông tin rút trích được từ corpora hoặc/và Wordnet [5], [6], [7], [8], [9]. Điểm giống nhau ở các nghiên cứu này là đa phần các phương pháp hiện tại có khả năng xây dựng được các tự điển có chất lượng cao (về cả số lượng entry và độ chính xác của chúng) cho các ngôn ngữ có sẵn nhiều nguồn tài nguyên từ vựng, hoặc phải sử dụng thêm các tài nguyên từ vựng ở nhiều ngôn ngữ trung gian. Độ chính xác của các từ điển song ngữ được xây dựng từ các từ điển sẵn có và Wordnet thường cao hơn so với sử dụng các tài nguyên từ vựng khác. Tuy nhiên, không phải tất cả các ngôn ngữ trong từ điển hiện có đều có Wordnet và chi phí để xây dựng Wordnet không hề nhỏ.

IV. NGÔN NGỮ VÀ CÁC TỪ ĐIỂN SONG NGỮ SẴN CÓ

Phương pháp chúng tôi giới thiệu để xây dựng từ điển song ngữ là tổng quát và có thể áp dụng cho mọi ngôn ngữ. Tuy nhiên, để tiện cho việc chứng minh tính đúng đắn của phương pháp, chúng tôi sẽ xây dựng từ điển cho các ngôn ngữ mà chúng tôi có chuyên gia sẵn sàng hỗ trợ. Cụ thể chúng tôi sẽ xây dựng từ điển song ngữ cho các ngôn ngữ Ả Rập, Assamese, Hindi và Việt. Trong quá trình trình bày, chúng tôi sẽ luân phiên sử dụng tên ngôn ngữ hoặc mã code của của các ngôn ngữ. Mã code ISO 693-3 của ngôn ngữ Ả Rập, Assamese, Hindi và Việt theo thứ tự là *arb, asm, hin* và *vie*.

Chúng tôi nghiên cứu các từ điển song ngữ sẵn có từ nhiều nguồn khác nhau và nhận thấy các từ điển được định dạng rất khác nhau. Việc rút trích và làm sạch thông tin từ các từ điển sẵn có mất rất nhiều công sức và thời gian. Chúng tôi sử dụng 4 từ điển song ngữ. Mỗi từ điển sẽ dịch các từ vựng giữa một ngôn ngữ nguồn mà chúng tôi lựa chọn và một từ hoặc cụm từ ở một ngôn ngữ trung gian giàu tài nguyên (trong trường hợp của chúng tôi là tiếng Anh với mã code ISO 693-3 là *eng*). Các từ điển chúng tôi sử dụng bao gồm:

Từ điển Ả Rập-Anh, *Dict(arb,eng)*, từ điển Anh-Hindi, *Dict(eng,hin)*, và tự điển Anh-Việt, *Dict(eng,vie)*, được cung cấp bởi Panlex[3].

Từ điển Assamese-Anh, *Dict(asm,eng)*, được tích hợp từ hai từ điển cung cấp bởi Xobdo[4] và Panlex.

Các tài nguyên từ điển sẵn có rất khác nhau về số lượng entry như trình bày ở Bảng 1.

Bảng 1. Số entry trong từ điển song ngữ hiện có

Từ điển	Số entry	Từ điển	Số entry
Dict(arb,eng)	53.194	Dict(eng,hin)	33.234
Dict(asm,eng)	76.634	Dict(eng,vie)	231.665

V. PHƯƠNG PHÁP

Trong phần này chúng tôi sẽ đề xuất phương pháp xây dựng từ điển mới, *Dict(A,C)*, từ 2 từ điển song ngữ sẵn có, *Dict(A,B)* và *Dict(B,C)*, với một ngôn ngữ chung B. Cụ thể, từ 4 từ điển song ngữ sẵn có *Dict(arb,eng)*, *Dict(asm,eng)*, *Dict(eng,hin)* và *Dict(eng,vie)*, chúng tôi sẽ xây dựng 4 từ điển *Dict(arb,hin)*, *Dict(arb,vie)*, *Dict(asm,hin)* và *Dict(asm,vie)*.

Số lượng entry trong các từ điển sẵn có rất khác nhau, như đã trình bày ở Bảng 1. Nếu số lượng entry trong từ điển sẵn có thấp sẽ dẫn đến số lượng entry trong từ điển mới cũng không cao. Do đó, trước khi xây dựng các từ điển song ngữ mới, làm tăng số lượng entry trong các từ điển sẵn có là rất cần thiết.

A. Làm tăng số entry trong từ điển sẵn có

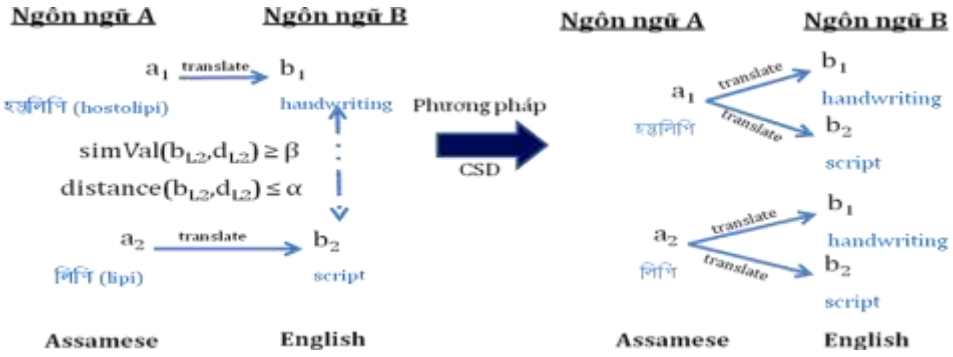
Lam và Kalita [5] giới thiệu các phương pháp xây dựng từ điển song ngữ mới có chiều dịch ngược với từ điển song ngữ hiện có và đồng thời làm tăng số lượng entry trong từ điển mới. Để làm tăng số entry trong từ điển, tác giả giới thiệu hai phương pháp DRwD và DRwS để tìm các từ hoặc cụm từ có nghĩa tương đương. Trong phương pháp DRwD, hai từ hoặc cụm từ được xem là có ngữ nghĩa tương đương nếu khoảng cách giữa chúng trong Princeton WordNet [10] nhỏ hơn ngưỡng α. Khoảng cách giữa hai từ trong Wordnet có giá trị từ 0,00 đến 1,00. Nếu hai từ hoặc

[3] http://panlex.org/
[4] http://www.xobdo.org/

cụm từ có khoảng cách là 0,00 thì có khả năng rất cao hai từ này có ngữ nghĩa giống nhau; ngược lại, nếu chúng có khoảng cách là 1,00 thì hai từ đó có ngữ nghĩa không giống nhau. Phương pháp DRwS cho phép tìm các từ và cụm từ có ngữ nghĩa giống nhau bằng cách tính giá trị *simVal* giữa các từ hoặc cụm từ. Nếu *simVal* của hai từ hoặc cụm từ càng lớn, thì khả năng chúng có ngữ nghĩa giống nhau càng cao. Giá trị *simVal* nằm trong khoảng từ 0,00 đến 1,00. S*imVal* giữa hai cụm từ là độ giống nhau giữa *ExpansionSet* của mỗi từ trong cụm từ. *ExpansionSet* của mỗi từ là tập giao của các "synset", "synonym", "hypernym" và "hyponym" của các từ đó trong WordNet.

Lam và Kalita kết luận phương pháp DRwS là phương pháp tốt nhất để tìm ra các từ đồng nghĩa. Tuy nhiên, trong quá trình thực nghiệm, chúng tôi phát hiện ra phương pháp DRwS vẫn còn một hạn chế có nguồn gốc từ chính Princeton Wordnet. Ví dụ, *simVal* của từ "mango" (nghĩa là "trái xoài" trong tiếng Việt) và "papaya" (nghĩa là "trái đu đủ" trong tiếng Việt) là 1,00 vì *ExpansionSet* của hai từ này là hoàn toàn giống nhau nên phương pháp DRwS kết luận "mango" và "papaya" có cùng ngữ nghĩa. Cụ thể hơn, từ hai entry ban đầu (mango, trái xoài) và (papaya, trái đu đủ), phương pháp DRwS tìm thêm 2 entry mới (mango, trái đu đủ) và (papaya, trái xoài). May mắn thay, khoảng cách giữa "mango" và "papaya" trong Princeton WordNet là 0,0769, do đó phương pháp DRwD với ngưỡng α là 0,00 sẽ kết luận "mango" khác với "papaya". Chúng tôi cũng phát hiện ra ExpansionSet của các số cũng giống nhau nên phương pháp DRwS cũng sẽ đưa ra những kết luận không chính xác như "sixteen" ("mười sáu") và "seventeen" ("mười bảy") có ngữ nghĩa giống nhau; trong khi đó phương pháp DRwD có thể đưa ra kết luận "sixteen" khác với "seventeen" do khoảng cách giữa chúng trong Wordnet là 0,125 (nếu đặt ngưỡng α là 0,00).

Để giải quyết vấn đề lỗi phát sinh từ Princeton Wordnet, chúng tôi kết hợp phương pháp DRwS và DRwD hình thành phương pháp CSD (Computing Similarity and Distance) để tìm ra các từ đồng nghĩa trong từ điển sẵn có. Một ví dụ khác minh họa ý tưởng của phương pháp CSD được trình bày trong Hình 1. Trong từ điển Assamese- Anh có 2 entry (hostolipi, handwriting) và (lipi, script). Từ từ điển Oxford English dictionary[5], "handwriting" nghĩa là "a particular form, style or method of writing by hand; the form or style of writing used by particular person" và "script" có nghĩa là "handwriting, the characters used in hand-writing (as distinguished from print)". Do đó, "handwriting" và "script" có nghĩa giống nhau. Phương pháp CSD cũng đưa ra kết luận là "handwriting" và "script" có nghĩa giống nhau. Như vậy, chúng ta tạo ra được 2 entry mới (হস্তলিপি, handwriting) và (লিপি, script) thêm vào từ điển Assamese-Anh.



Hình 1: Phương pháp CSD

Phương pháp CSD được trình bày trong Giải thuật 1. Xét 2 *LexicalEntry* có cùng thông tin về loại từ POS (Giải thuật 1, dòng 1-4), nếu giá trị *simVal* của $LexicalEntry_i$ và $LexicalEntry_j$ lớn hơn hoặc bằng một ngưỡng β (Giải thuật 1, dòng 5) và khoảng cách giữa $LexicalEntry_i$ và $LexicalEntry_j$ nhỏ hơn hoặc bằng một ngưỡng α (Giải thuật 1, dòng 6), phương pháp CSD sẽ kết luận là 2 *LexicalEntry* này có ngữ nghĩa giống nhau và thêm entry mới tìm vào từ điển (Giải thuật 1, dòng 7).

Giải thuật 1: Phương pháp CSD

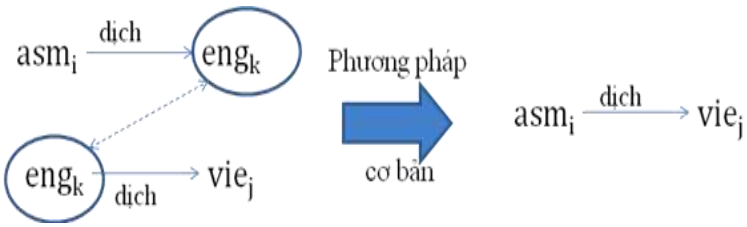
1: for all $LexicalEntry_i$
2: for all $Sense_u \in LexicalEntry_i$
3: for all $LexicalEntry_j$ having the same POS with $LexicalEntry_i$ do
4: for all $Sense_v \in LexicalEntry_j$ do
5: if $simVal(LexicalEntry_i, LexicalEntry_j) \geq \beta$ then
6: if $distance(LexicalEntry_i, LexicalEntry_j) \leq \alpha$ then
7: add <$LexicalEntry_i$.LexicalUnit, $Sense_v$> to Dictionary
8: end if

[5] http://www.oed.com/

```
9:            end if
10:         end for
11:       end for
12:     end for
13: end for
```

B. Phương pháp cơ bản (baseline approach) để xây dựng từ điển mới

Sau khi làm tăng số entry trong từ điển, chúng tôi bắt đầu xây dựng các từ điển mới. Đầu tiên chúng tôi giới thiệu phương pháp cơ bản để xây dựng một từ điển mới từ các từ điển sẵn có. Ví dụ minh họa cho phương pháp cơ bản được trình bày trong Hình 2. Cho từ điển song ngữ Assamese-Anh chứa các entry (asm_i, eng_k) và một tự điển Anh-Việt chứa các entry (eng_k, vie_j), chúng tôi sẽ xây dựng từ điển mới Assamese-Việt chứa entry (asm_i, vie_j) nếu cả asm_i và vie_j đều có mối quan hệ với eng_k.



Hình 2. Phương pháp cơ bản để xây dựng từ điển song ngữ

Phương pháp cơ bản để xây dựng một từ điển mới từ 2 từ điển sẵn có được trình bày trong Giải thuật 2. Chúng ta xây dựng từ điển mới từ hai từ điển nguồn *Dict(A,B)* và *Dict(B,C)*. Với mỗi $LexicalEntry_i$ trong *Dict(A,B)* và mỗi $LexicalEntry_j$ trong *Dict(B,C)* có cùng thông tin POS (Giải thuật 2, dòng 1-3), nếu tồn tại $LexicalEntry_i.Sense$ giống $LexicalEntry_j.LexicalUnit$ (Giải thuật 2, dòng 4) thì ta thêm $<LexicalEntry_i.LexicalUnit, LexicalEntry_j.Sense>$ vào từ điển mới *Dict(A,C)* (Giải thuật 2, dòng 5).

Giải thuật 2: Phương pháp cơ bản

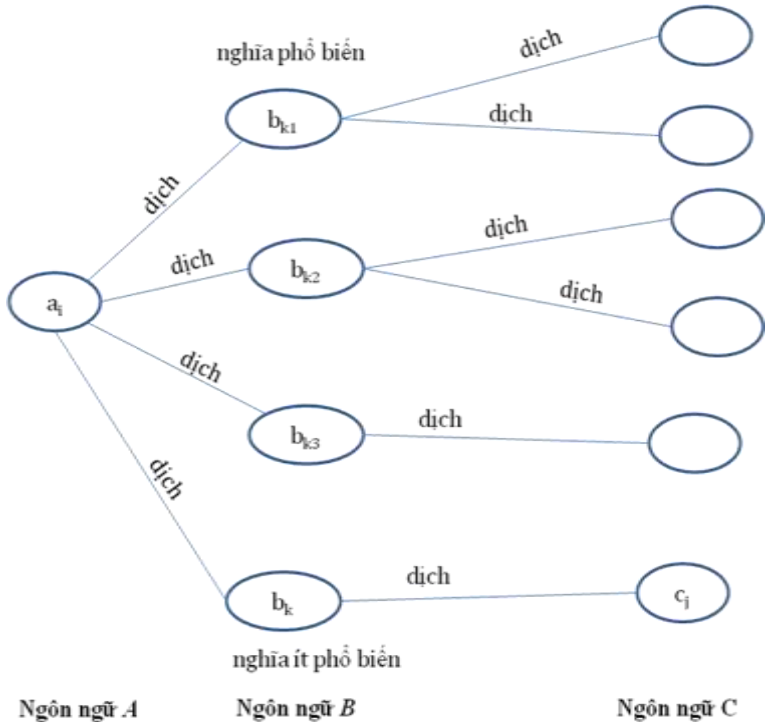
```
Input: Dict(A,B) và Dict(B,C)
Output: Dict(A,C)

1: Dict(A,C): =Ø
2: for all LexicalEntry_i ∈ Dict(A,B) do
3:       for all LexicalEntry_j ∈ Dict(B,C) having the same POS with LexicalEntry_i do
4:             if LexicalEntry_i.Sense = LexicalEntry_j.LexicalUnit then
5:                   add <LexicalEntry_i.LexicalUnit, LexicalEntry_j.Sense> to Dict(A,C)
6:             end if
7:       end for
8: end for
```

C. Phương pháp ánh xạ trung gian

Phương pháp cơ bản có ưu điểm là xây dựng một từ điển mới rất nhanh, chi phí thấp. Tuy nhiên, nếu từ ở ngôn ngữ trung gian có nhiều hơn một nghĩa hay đa nghĩa thì phương pháp cơ bản có khả năng sẽ đưa ra những kết luận sai. Cụ thể hơn, nếu b_k có hai nghĩa và được dịch sang ngôn ngữ C tương ứng là c_{j1} và c_{j2}, phương pháp cơ bản sẽ kết luận từ a_i ở ngôn ngữ A có hai nghĩa ở ngôn ngữ C và thêm hai entry (a_i, c_{j1}) và (a_i, c_{j2}) vào từ điển *Dict(A,C)*, điều này chưa chắc đúng. Để giảm bớt những entry không chính xác trong từ điển mới, chúng tôi sử dụng phương pháp tính trọng số entry dựa vào tính phổ biến trong ngữ nghĩa của từ ở ngôn ngữ trung gian.

Trong một từ điển, các *sense* ở ngôn ngữ đích thường được sắp xếp theo thứ tự dựa vào tính phổ biến nghĩa của từ. Với mỗi *LexicalUnit*, *sense* đầu tiên thường có tính phổ biến nhất trong ngôn ngữ đích, và ngược lại cho *sense* cuối cùng. Giả sử b_k là một *sense* hiếm khi được sử dụng của a_i và do b_k chỉ có một nghĩa c_j ở ngôn ngữ C, hầu hết các phương pháp hiện tại đều kết luận rằng a_i được dịch sang c_j. Sự thật thì đây là một entry kém chất lượng vì mối quan hệ giữa a_i và b_k rất yếu. Ví dụ minh họa được thể hiện trong Hình 3.



Hình 3. Liên kết yếu (a_i,b_k) trong từ điển

Để giảm bớt các entry kém chất lượng vì mối quan hệ yếu giữa a_i và b_k và/hoặc giữa b_k và c_j, đầu tiên chúng tôi tính trọng lượng, còn gọi là *weight*, cho mỗi *sense* của từng *LexicalUnit* dựa vào độ phổ biến của chúng. *Sense* có độ phổ biến cao hơn sẽ có *weight* lớn hơn và ngược lại. Phương pháp tính *weight* cho mỗi *sense* được trình bày ở Giải thuật 3.

Giải thuật 3: Phương pháp tính *weight* cho mỗi *sense* của mỗi *LexicalUnit*

1: t_tr ← total translations of a_i
2: temp := 0
3: for all translations b_j of a_i do
4: $temp$ += $rank_{bj}$
5: end for
6: for all translations b_j of a_i do
7: $weight(a_i,b_j) = \frac{t_{tr} - rank_{bj} + 1}{temp}$
8: end for

Trong *Dict(A,B)*, giả sử từ a_i có bốn *sense* theo thứ tự là b_1, b_2, b_3 và b_4. Giá trị *rank* dựa trên độ phổ biến của các *sense* được trình bày trong Bảng 2. Sense b_1 có độ phổ biến nhất nên *rank* là 1 và b_4 tương ứng sẽ có *rank* là 4. Để dễ hiểu, chúng tôi tách các *sense* của từ và hình thành một *entry* với *rank* tương ứng được trình bày trong Bảng 2.

Bảng 2. Ví dụ một *LexicalUnit* có 4 *sense*

Entry	Rank	Entry	Rank
(a_i,b_1)	1	(a_i,b_3)	3
(a_i,b_2)	2	(a_i,b_4)	4

Tổng số *sense* của a_i trong ngôn ngữ B, được gọi là *t_tr*, là 4 (Giải thuật 3, dòng 1). Một giá trị tạm thời *temp* (Giải thuật 3, dòng 2) là tổng số *rank* của các entry:

$$temp = 1 + 2 + 3 + 4 = 10$$

Do đó, *weight* của mỗi entry được tính như sau (Giải thuật 3, dòng 7):

weight(a_i,b_1)=(4-1+1)/10=0.4 weight(a_i,b_2)=(4-2+1)/10=0.3

weight(a_i,b_3)=(4-3+1)/10=0.2 weight(a_i,b_4)=(4-4+1)/10=0.1

Thực hiện tương tự để tính *weight* cho các entry trong từ điển còn lại, *Dict(B,C)*. Giả sử tồn tại entry (a_i,b_k) trong *Dict(A,B)* và entry (b_k,c_j) trong *Dict(B,C)* có quan hệ với nhau thông qua từ b_k ở ngôn ngữ trung gian B. Theo phương pháp cơ bản, chúng ta sẽ có entry tiềm năng (a_i,c_j) trong từ điển *Dict(A,C)*. Tiếp theo, chúng tôi tính giá trị *score* của entry tiềm năng (a_i,c_j), hay còn gọi là $score(a_i,c_j)$. Nếu $score(a_i,c_j)$ lớn hơn một ngưỡng δ, chúng tôi kết luận đây là một entry tốt và chèn nó vào *Dict(A,C)*. $Score(a_i,c_j)$ được tính là tích của $weight(a_i,b_k)$ và $weight(b_k,c_j)$:

$$score(a_i, c_j) = weight(a_i, b_k) \times weight(b_k, c_j)$$

VI. KẾT QUẢ THỰC NGHIỆM

A. Chuẩn hóa dữ liệu

Trước khi thực hiện xây dựng từ điển mới, chúng tôi cần tiến hành chuẩn hóa dữ liệu. Đầu tiên, chúng tôi cần loại bỏ các từ nằm trong danh sách "stop words[6]" như "someone", "to" và "that". Sau đó, thực hiện chuẩn hóa các từ hoặc cụm từ về từ gốc của chúng (steam word). Chẳng hạn, chuẩn hóa từ "teaching" thành "teach". Phương pháp nổi tiếng để chuẩn hóa các từ tiếng Anh là phương pháp Porter stemmer [11]. Tuy nhiên, chúng tôi không thể sử dụng phương pháp này vì một số trường hợp từ sau khi chuẩn hóa không có nghĩa. Ví dụ, Porter stemmer chuẩn hóa từ "imitate", "language" và "software" thành các từ không có nghĩa "imit", "languag" và "softwar". Thêm vào đó, do chúng tôi cần tìm *ExpansionSet* bao gồm các synset, synonym, hypernym và hyponym của các từ tiếng Anh từ Princeton Wordnet để tính toán độ giống nhau về mặt ngữ nghĩa của các từ trong từ điển. Do đó, chúng tôi sử dụng hàm chuẩn hóa từ do Rita.Wordnet[7] cung cấp. Mặc dù hàm chuẩn hóa của Rita.Wordnet cũng không chính xác hoàn toàn, nhưng Rita.Wordnet cung cấp các hàm hỗ trợ tìm *ExpansionSet* cho các từ do Rita.Wordnet chuẩn hóa. Vì vậy, việc chuẩn hóa từ bằng Rita.Wordnet vẫn chấp nhận được.

Thông tin về POS của mỗi entry trong từ điển đóng vai trò rất quan trọng trong việc tìm ra các từ hoặc cụm từ có nghĩa tương đương từ Wordnet hoặc xây dựng các entry tiềm năng trong từ điển mới. Tuy nhiên, không phải mọi entry trong từ điển đều chứa thông tin POS. Cụ thể, 100% entry trong tự điển Ả Rập-Anh và 6,63% entry trong từ điển Anh-Việt không chứa thông tin POS. Để tìm POS cho các entry không có thông tin POS, chúng tôi sử dụng thông tin POS phổ biến nhất (the best POS) của từ tiếng Anh trong mỗi entry. Thông tin POS phổ biến nhất của từ được cung cấp bởi Rita.Wordnet.

B. Phương pháp đánh giá

Phương pháp tiêu chuẩn để đánh giá một từ điển song ngữ do máy xây dựng là yêu cầu người dùng đánh giá toàn bộ các entry trong từ điển đó. Một điểm cần lưu ý là các người dùng phải sử dụng thành thạo cả ngôn ngữ nguồn và ngôn ngữ đích trong mỗi từ điển họ tham gia đánh giá. Tuy nhiên, để tìm ra những người dùng thành thạo, nắm được mọi ngữ nghĩa của tất cả các từ ở cả 2 ngôn ngữ trong một từ điển song ngữ không phải là chuyện đơn giản. Thêm vào đó, do một trong những mục tiêu của chúng tôi là xây dựng từ điển cho các ngôn ngữ có ít tài nguyên (Assamese), chúng tôi không thể tìm ra bất cứ người dùng nào có thể thành thạo cả 2 ngôn ngữ trong những từ điển: Assamese-Việt, Arabic-Việt, Arabic-Hindi. Vì vậy, cho mỗi từ điển mới cần đánh giá, chúng tôi nhờ các cặp người dùng đánh giá. Trong mỗi cặp đánh giá, mỗi người dùng thành thạo một ngôn ngữ trong từ điển và một ngôn ngữ trung gian. Hai người dùng này sẽ giao tiếp thông qua ngôn ngữ trung gian (tiếng Anh) để đánh giá các entry trong từ điển. Riêng từ điển Assamese-Hindi được đánh giá bằng những người dùng thành thạo cả hai ngôn ngữ.

Đánh giá toàn bộ entry trong một từ điển sẽ tốn rất nhiều thời gian. Dựa vào qui luật "general rules of thumb" [12], chúng tôi có thể chọn ngẫu nhiên 30 entry trong từ điển và yêu cầu người dùng đánh giá. Để đảm bảo độ chính xác cao nhất có thể, chúng tôi chọn ngẫu nhiên 100 entry trong mỗi từ điển và yêu cầu 4-5 người dùng (hoặc cặp người dùng) đánh giá sử dụng thang 5-điểm: 5: rất chính xác (Excellent), 4: tốt (Good), 3: trung bình (Average), 2: tạm chấp nhận (Fair) và 1: sai (Bad).

C. Kết quả

Để đánh giá được sự ảnh hưởng của chất lượng các từ điển sẵn có đến chất lượng các từ điển mới, chúng tôi cũng tiến hành đánh giá 4 từ điển mà chúng tôi sử dụng như các tài nguyên đầu vào. Bảng 3 trình bày điểm trung bình của các entry trong từ điển sẵn có. Mức độ đồng ý giữa những người đánh giá là khoảng 70%.

Bảng 3. Điểm trung bình của các entry trong từ điển sẵn có

Từ điển	Điểm	Từ điển	Điểm
Dict(arb,eng)	3,58	Dict(eng,hin)	3,70
Dict(asm,eng)	4,65	Dict(eng,vie)	3,77

[6] http://www.world-english.org/english500.htm
[7] http://rednoise.org/rita/index.html

Theo Lam và Kalita [5], phương pháp tốt nhất để tìm ra các entry mới trong từ điển sẵn có là phương pháp DRwS. Để chứng minh là cần phải kết hợp cả tìm độ giống nhau giữa các *ExpansionSet* của từ và khoảng cách của từ trong Wordnet, chúng tôi tiến hành thực nghiệm cả hai phương pháp DRwS và CSD và tiến hành đánh giá, so sánh. Điểm trung bình và số lượng các entry mới được tạo ra khi sử dụng phương pháp DRwS và CSD để tìm ra các entry mới trong từ điển sẵn có được trình bày lần lượt trong Bảng 4 và Bảng 5.

Bảng 4. Điểm trung bình và số lượng các entry mới được tạo ra bằng phương pháp DRwS

Từ điển	DRwS ($\beta \geq 0{,}90$)		DRwS ($\beta = 1{,}00$)	
	Điểm	Entry mới	Điểm	Entry mới
Dict(arb,eng)	1,62	19.547	1,70	15.621
Dict(asm,eng)	2,67	11.548	4,01	8.581
Dict(eng,hin)	3,30	7.125	3,60	3.120
Dict (eng,vie)	2,01	58.446	3,14	28.532

Bảng 5. Điểm trung bình và số lượng các entry mới được tạo ra bằng phương pháp CSD

Từ điển	CSD ($\beta \geq 0{,}90$ & $\alpha = 0{,}00$)		CSD ($\beta = 1{,}00$ & $\alpha = 0{,}00$)	
	Điểm	Entry mới	Điểm	Entry mới
Dict(arb,eng)	2,93	10.189	2,68	7.120
Dict(asm,eng)	4,20	1.120	4,31	530
Dict(eng,hin)	3,38	5.623	3.67	840
Dict (eng,vie)	3,51	36.124	3,58	10.123

Phương pháp CSD tìm ra ít entry mới hơn phương pháp DRwS; tuy nhiên, độ chính xác của các entry mới tạo bằng phương pháp CSD là cao hơn phương pháp DRwS. Chúng tôi chỉ thêm các entry mới xây dựng bằng phương pháp CSD với ngưỡng $\beta = 1{,}00$ và $\alpha = 0{,}00$ vào từ điển.

Sau khi làm tăng số entry trong từ điển sẵn có, chúng tôi tiến hành xây dựng từ điển mới bằng phương pháp cơ bản và phương pháp ánh xạ trung gian. Điểm trung bình và số lượng các entry trong từ điển mới được trình bày trong Bảng 6. Phương pháp ánh xạ trung gian kết hợp với tính *score* của các entry tiềm năng làm giảm số lượng của các entry kém chất lượng có trong từ điển so với phương pháp cơ bản. Từ thực nghiệm, nếu δ là 0.40 sẽ giúp tạo ra các từ điển có chất lượng tốt nhất, tuy nhiên số lượng entry trong từ điển không cao.

Bảng 6. Điểm trung bình và số lượng entry trong từ điển mới xây dựng

Phương pháp cơ bản			Phương pháp ánh xạ trung gian ($\delta \geq 0.1$)		
Từ điển	**Điểm**	**Entry**	**Từ điển**	**Điểm**	**Entry**
Dict(arb,vie)	2,06	270.048	Dict(arb,vie)	2,15	84.048
Dict(asm,vie)	3,00	308.129	Dict(asm,vie)	3,40	108.129
Dict(arb, hin)	2,34	140.153	Dict(arb, hin)	2,61	50.153
Dict (asm, hin)	2,50	102.138	Dict (asm, hin)	3,50	42.138
Phương pháp ánh xạ trung gian ($\delta \geq 0.2$)			**Phương pháp ánh xạ trung gian ($\delta \geq 0.4$)**		
Từ điển	**Điểm**	**Entry**	**Từ điển**	**Điểm**	**Entry**
Dict(arb,vie)	3,23	28.965	Dict(arb,vie)	3,60	12.129
Dict(asm,vie)	3,55	40.220	Dict(asm,vie)	3,89	23.248
Dict(arb, hin)	3,45	15.864	Dict(arb, hin)	3,68	9.196
Dict (asm, hin)	3,69	13.127	Dict (asm, hin)	4,01	8.349

D. Thảo luận

Các phương pháp làm tăng số lượng entry trong từ điển (DRwD, DRwS và CSD), phương pháp cơ bản và phương pháp ánh xạ trung gian để xây dựng từ điển mới đều phải sử dụng thông tin POS trong mỗi entry. Nếu một từ điển sẵn có chứa đầy đủ thông tin POS thì các entry mới tạo có độ chính xác cao; và ngược lại. Ví dụ, từ điển Ả Rập-Anh hoàn toàn không chứa thông tin POS nên độ chính xác của các entry mới rất thấp so với các entry mới tạo từ các từ điển có chứa đầy đủ thông tin POS như từ điển Assamese-Anh. Thực tế thì một số ngôn ngữ có rất ít từ điển và số từ điển hiện có này chỉ chứ các từ hoặc cụm từ ở ngôn ngữ nguồn và các nghĩa tương ứng ở ngôn ngữ đích, hoàn toàn không chứa bất cứ thông tin nào khác như POS hay các ví dụ minh họa cách sử dụng từ. Nghiên cứu các giải pháp để tìm thông tin POS cho các entry trong một từ điển sẵn có rất đáng quan tâm. Hiện tại chúng tôi chỉ gán thông tin POS phổ biến nhất của từ tiếng Anh cho entry không có POS và cách làm này có khả năng không chính xác. Ví dụ, từ "book" có thể là danh từ "noun" hoặc là động từ "verb". Do POS phổ biến nhất của "book" là "noun", nên tất cả các

entry trong từ điển không có POS mà có nghĩa là "book" đều được gán POS là "noun". Kết quả là rất nhiều entry của "book" có POS là "verb" sẽ có độ chính xác không cao.

Sau khi tìm ra được các entry mới thì việc sắp xếp các nghĩa theo mức độ phổ biến của chúng trong thực tế cũng rất quan trọng. Cụ thể, phương pháp ánh xạ trung gian mà chúng tôi giới thiệu cần thông tin độ phổ biến của ngữ nghĩa để loại bỏ bớt các entry kém chất lượng. Ví dụ , trong trong Assamese từ " আৰক্ষণ কৰ্ " có POS là "verb" và nghĩa là "book". Sau khi áp dung phương pháp CSD tìm thêm từ đồng nghĩa thì "আৰক্ষণ কৰ্ " với POS "verb" có 2 nghĩa "book" và "reserve". Vậy giữa "book" và "reserve", từ nào có mức độ phổ biến hơn trong thực tế? Hiện tại chúng tôi chỉ mới tìm ra được các entry mới, còn việc sắp xếp các entry theo mức độ phổ biến thì cần phải có nhiều tài nguyên hơn, chẳng hạn như các tài liệu văn bản ở ngôn ngữ nguồn, ngôn ngữ đích hoặc các tài liệu song ngữ.

Chúng tôi chỉ mới tìm hiểu phương pháp xây dựng từ điển mới *Dict(A,C)* từ *Dict(A,B)* và *Dict(B,C)*. Nếu thay đổi chiều các từ điển sẵn có để xây dựng từ điển mới thì kết quả có ảnh hưởng như thế nào? Ví dụ, nếu chúng ta sử dụng *Dict(A,B)* và *Dict(C,B)* để xây dựng từ điển *Dict(A,C)* hoặc *Dict(C,A)* thì kết quả có tối ưu hơn hay không? Hoặc nếu chúng ta sử dụng *Dict(B,A)* và *Dict(B,C)* để xây dựng *Dict(A,C)* hoặc *Dict(C,A)* thì kết quả sẽ có gì khác biệt? Chiều trong từ điển song ngữ sẽ ảnh hưởng đến số lượng entry và độ chính xác của các entry ra sao vẫn là câu hỏi cần nghiên cứu sâu hơn.

Trong các từ điển hiện có có chứa nhiều từ có nghĩa hiếm khi được sử dụng hoặc chứa các từ cổ. Chúng tôi nhận thấy người đánh giá thường cho điểm rất thấp cho các từ nằm trong dạng hiếm sử dụng hoặc từ cổ. Bên cạnh đó, bản thân từ điển sẵn có cũng chứa đựng các entry mà người dùng không biết. Thêm vào đó, chắc chắn độ chính xác của các entry mới tìm sẽ phụ thuộc rất lớn và độ chính xác của các entry trong từ điển sẵn có. Nếu từ điển sẵn có chứa các entry không chính xác thì entry mới tìm được cũng sẽ có độ chính xác không cao. Bảng 7 trình bày một số entry trong từ điển sẵn có mà người dùng không biết và đánh giá điểm thấp. Bảng 8 trình bày một số entry không chính xác trong từ điển sẵn có.

Bảng 7. Một số từ và cụm từ người dùng không biết

Arabic word	**Evaluation**	**Note**
إيكيلون	Bad	Do not know *arb* word
خطآ	Bad	Do not know *arb* word
خواص غروانية	Bad	Do not know *arb* word
Assamese word	**Evaluation**	**Note**
অৱচে;লাই	Bad	Do not know *asm* word
অপ্রভুল	Bad	Do not know *asm* word
ইন্দ্রবস্তি	bad	Do not know *asm* word
Vietnamese word	**Evaluation**	**Note**
báo cừu	Bad	Do not know *vie* word
bì xì	Bad	Do not know *vie* word
diện địa	Bad	Do not know *vie* word

Bảng 8. Một số entry không chính xác trong tự điển sẵn có

Arabic word	**POS**	**English word**	**Evaluation**	**Note**
زوج	NULL	manacles	Bad	The correct meaning of the *arb* word is "couple"
جاي	NULL	gay	Bad	Using *arb* language to write the *eng* word
صح	NULL	health	Bad	The correct meaning of the *arb* word is "true"

Assamese word	POS	English word	Evaluation	Note
নেওচা	n	curse	Bad	The correct meaning of the *asm* word is “ignore”
কপিঞ্জল	n	skylark	Bad	The correct meaning of the *asm* word is “sky”
অভিনয়	n	cast	Good	The correct meaning of the *asm* word is “acting”
শিলিখা	n	haritaki	Bad	Do not know the *eng* word “haritaki”
কুবিয়া	n	strike	Fair	Not good spelling in *asm* word
Vietnamese word	**POS**	**English word**	**Evaluation**	**Note**
luôn	NULL	sempre	Bad	The *vie* word should combine with other vie words to create a real compound word. The eng word is not known.
La	n	tuberculosis	Average	The *vie* word should combine with another word such as “bệnh” or “bịnh” to create “bệnh lao” or “bịnh lao” having the meaning of “tuberculosis”
kỹ thuật	NULL	techie	Fair	The correct meaning of the *vie* word is “technology”

VII. KẾT LUẬN

Mục đích của nghiên cứu này là xây dựng từ điển mới cho các ngôn ngữ không có nhiều nguồn tài nguyên từ vựng. Chúng tôi đã làm tăng số lượng entry trong từ điển, đã có thể xác định và loại bỏ được các entry hiếm hoặc entry có chất lượng không tốt trong từ điển mới. Bước kế tiếp, chúng tôi sẽ cải tiến giải thuật để xây dựng các từ điển mới có chất lượng tốt hơn và số entry nhiều hơn. Bên cạnh đó, chúng tôi sẽ sử dụng các nguồn tài nguyên sẵn có ở các ngôn ngữ trung gian khác để làm tăng số entry trong từ điển mới chẳng hạn như sử dụng Wordnet làm tài nguyên trung gian [13].

VIII. LỜI CẢM ƠN

Chúng tôi xin chân thành cảm ơn sự hỗ trợ của các bạn trong dự án Panlex và Xobdo đã cung cấp các từ điển song ngữ cho chúng tôi nghiên cứu. Chúng tôi rất cám ơn sự giúp đỡ nhiệt tình của Jugal Kalita, Dubari Borah, Tri Doan, Abhijit Bendale, Lalit Prithviraj Jain, Svati Dhamija, Hoang Nguyen, Cuong Nguyen, Bai Le, Feras Al. Tarouti và Faris Kateb trong việc hỗ trợ đánh giá các từ điển.

TÀI LIỆU THAM KHẢO

[1] S. I. Landau, Dictionaries: The art and craft of lexicography, Macmillan Reference USA, 1984.

[2] Kumiko Tanaka and Kyoji Umemura, "Construction of a bilingual dictionary intermediated," in *Proceedings of the 15th Conference on Computational Linguistics (COLING), volume 1*, Kyoto, Japan, 1994.

[3] Tim Gollins and Mark Sanderson, "Improving cross language information retrieval with triangulated translation," in *Proceedings of the 24th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, New York, USA, 2001.

[4] Kisuh Ahn and Matthew Frampton, "Automatic generation of translation dictionaries," in *Proceedings of the International Workshop on CrossLanguage Knowledge Induction*, Trento, Italy, 2006.

[5] Khang Nhut Lam and Jugal Kalita, "Creating reverse bilingual dictionaries," in *The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, USA, 2013.

[6] R. D. Brown, "Automated dictionary extraction for "Knowledge-free" example-based translation," in *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation*, Santa Fe, USA, 1997.

[7] Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsumoto, "Building a Japanese-Chinese dictionary using Kanji/Hanzi conversion," in *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP)*, Jeju Island, Korea, 2005.

[8] Nikola Ljube and Darja Fiser, "Bootstrapping bilingual lexicons from comparable corpora for closely related languages," in *Proceedings of the 14th International Conference on Text, Speech and Dialogue (TSD)*, Plzen, Czech Republic, 2011.

[9] Pablo G. Otero and Jose R.P. Campos, "Automatic generation of bilingual dictionaries using intermediate languages and comparable corpora," in *2010*, Romania, in *Proceedings of the 11th International Conference on Computational Linguistics and Intelligent Text Processing* (CICLing).

[10] G. Miller, "Wordnet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995.

[11] M. F. Porter, "An algorithm for suffix stripping," *Program: Electronic library and information system*, vol. 3, no. 40, pp. 211-218, 2006.

[12] S. M. Ross, Introductory statistics, 2010: Academic Press.

[13] Khang Nhut Lam, Feras Al Tarouti, and Jugal K. Kalita, "Automatically Creating a Large Number of New Bilingual Dictionaries," in *AAAI*, Texas, USA, 2015.

[14] G. G. Koch, Intraclass correlation coefficient. Encyclopedia of statistical sciences, John Wiley & Sons, 1982.

CONSTRUCTING BILINGUAL DICTIONARIES USING TRANSITIVITY

Khang Nhut Lam

ABSTRACT*— To construct a bilingual dictionary from a source language A to a target language C, the so-called Dict(A,C), existing approaches need many existing lexical resources such as intermediate dictionaries or corpora in A, C and other intermediate languages. However, not all of languages have these resources, specially resource poor and endangered languages reported by UNESCO. This paper presents approaches to increase the number of entries in an existing dictionary and to create new bilingual dictionaries from existing bilingual dictionaries Dict(A,C) from Dict(A,B) and Dict(B,C) using transitivity. To handle ambiguity, we introduce a weighting scheme method such that if an entry has a weighting score greater than a threshold δ, we accept it as a correct translation and add it to the new dictionary. Our research helps not only reduce the cost to construct new bilingual dictionaries but also support communities using resource poor languages.*