PHƯƠNG PHÁP VÀ CÔNG CỤ ĐÁNH GIÁ TỰ ĐỘNG CÁC HỆ THỐNG DỊCH TỰ ĐỘNG TRÊN MẠNG

METHODS AND TOOL FOR THE AUTOMATIC EVALUATION OF FREE ONLINE TRANSLATORS

VÕ TRUNG HÙNG
Trường Đại học Bách Khoa, Đại học Đà Nẵng

TÓM TẮT
Trong bài báo này chúng tôi giới thiệu những phương pháp cho phép đánh giá chất lượng của một bản dịch theo phương pháp NIST và BLEU. Tiếp theo, chúng tôi giới thiệu công cụ do chúng tôi phát triển để đánh giá tự động chất lượng của các hệ thống dịch tự động trên mạng như Reverso, Sytran...

ABSTRACT
In this paper, we introduce methods to evaluate quality of a translation by NIST and BLEU. We present also a tools that we was developed for automatic evaluation quality of free online translators such as Reverso, Sytran...

1 Giới thiệu

Hiện tại, chúng ta có thể tìm thấy ngày càng nhiều những hệ thống dịch tự động miễn phí trên mạng như: Systran, Reverso, WorldLingo, IBM translator… Những hệ thống này cho phép dịch tự động các văn bản với một cặp ngôn ngữ chọn trước (ví dụ: dịch một văn bản tiếng Anh sang tiếng Pháp).

Tuy nhiên, chất lượng dịch là vấn đề mà người sử dụng quan tâm vì đa số các hệ thống dịch tự động hiện nay có chất lượng khá thấp. Để dịch một văn bản từ tiếng Anh sang tiếng Pháp chúng ta có thể chọn sử dụng nhiều hệ thống dịch khác nhau và kết quả nhận được cũng sẽ (có thể) khác nhau. Vấn đề đặt ra là người sử dụng nên chọn sử dụng hệ thống dịch nào cho văn bản của mình ?
Trong khuôn khổ dự án TraCorpEx, đây là dự án hợp tác giữa Trung tâm Nghiên cứu Ứng dụng Công nghệ Thông tin và Truyền thông (DATIC, Trường Đại học Bách Khoa, Đại học Đà Nẵng) với Trung tâm nghiên cứu GETA (Trung tâm nghiên cứu dịch tự động và xử lý ngôn ngữ tự nhiên, Cộng hoà Pháp) về dịch tự động trong việc sử dụng kết hợp nhiều hệ thống dịch khác nhau; chúng tôi đã nghiên cứu và phát triển một công cụ cho phép đánh giá tự động chất lượng của một vài hệ thống dịch tự động trên cơ sở sử dụng phương pháp BLEU (BiLingual Evaluation Understudy) và NIST (National Institute of Standards and Technology).

Với công cụ này, chúng ta có thể đánh giá chất lượng của một hệ thống dịch tự động thông qua một kho dữ liệu (corpus) gồm các câu gốc và các câu dịch tham khảo. Công cụ của chúng tôi cho phép xử lý và gửi các câu của một văn bản gốc đến các hệ thống dịch, tiếp theo lấy kết quả nhận được sau khi dịch đối chiếu với dữ liệu tham khảo (thông thường là các bản dịch chuẩn) để tính điểm phục vụ việc đánh giá [9]. Công cụ này có thể thực hiện được trên Internet hoặc trên máy đơn.

Trong bài báo này, chúng tôi trình bày những phương pháp đánh giá chất lượng bản dịch và cách xây dựng một công cụ dựa trên các phương pháp đó. Đồng thời, chúng tôi cũng đưa ra một số kết quả thử nghiệm trên cơ sở đánh giá hai hệ thống dịch được sử dụng phổ biến hiện nay là Systran, Reverso trên các dữ liệu có sẵn của BTEC, BIBLE.

2 Nghiên cứu tổng quan

Trong phần này chúng tôi giới thiệu vắn tắt một số hệ thống dịch tự động đang được sử dụng rộng rãi hiện nay và các phương pháp để đánh giá chất lượng bản dịch.

2.1 Các hệ thống dịch tự động

2.1.1 Systran

Hiện tại, SYSTRAN là một hệ thống dịch tự động rất nổi tiếng và chất lượng dịch khá tốt. SYTRAN có thể sử dụng được trên môi trương Internet, máy đơn hoặc trên các hệ thống mạng cục bộ. Nó có thể dịch được cho 36 cặp ngôn ngữ và người dùng có thể chọn dịch các văn bản chuyên ngành cho 20 lĩnh vực khác nhau. Phiên bản dùng trên Internet có thể dịch cho 34 cặp ngôn ngữ và đặt tại địa chỉ: http://www.systranbox.com/.

2.1.2 Gist-In-Time

Đây là hệ thống dịch được phát triển bởi Alis Technologies Inc. Nó có thể dịch cho 17 cặp ngôn ngữ. Để có thể truy cập vào hệ thống dịch Gist-In-Time trên Internet chúng ta có thể truy cập vào: http://www.teletranslator.com:8100/cgi-bin/transint.fr.pl?AlisTargetHost=localhost

2.1.3 Reverso

Đây là hệ thống dịch tự động của Softissimo để dịch các văn bản hoặc các trang Web dưới dạng HTML. Hệ thống này có thể thực hiện được trên Internet, Intranet hoặc như là một ứng dụng độc lập trên máy đơn. Địa chỉ của hệ thống dịch trên Internet là: http://www.reverso.net/textonly/default.asp.

2.2 Các phương pháp đánh giá các bản dịch

Trong phần này chúng tôi giới thiệu vắn tắt hai phương pháp được sử dụng để đánh giá những bản dịch: BLEU và NIST. Những phương pháp này dựa trên cơ sở đánh giá mức độ trùng khớp các dãy ký tự có độ dài n (phương pháp n-grams) giữa bản dịch bằng máy và bản dịch tham khảo để đánh giá [7] và [9].

2.2.1 BLEU

BLEU là một phương pháp dùng để đánh giá chất lượng bản dịch được đề xuất bởi IBM tại hội nghị ACL ở Philadelphie vào tháng 7-2001 [6]. Ý tưởng chính của phương pháp là so sánh kết quả bản dịch tự động bằng máy với một bản dịch chuẩn dùng làm bản đối chiếu. Việc so sánh được thực hiện thông qua việc thống kê sự trùng khớp của các từ trong hai bản dịch có tính đến thứ tự của chúng trong câu (phương pháp n-grams theo từ) [3]. Phương pháp này dựa trên hệ số tương quan giữa bản dịch máy và bản dịch chính xác được thực hiện bởi con người để đánh giá chất lượng của một hệ thống dịch.

Việc đánh giá được thực hiện trên kết quả thống kê mức độ trùng khớp các n-grams (dãy ký tự gồm n từ hoặc ký tự) từ kho dữ liệu của kết quả dịch và kho các bản dịch tham khảo có chất lượng cao [5]. Giải thuật của IBM đánh giá chất lượng của hệ thống dịch qua việc trùng khớp của các n-grams đồng thời nó cũng dựa trên cả việc so sánh độ dài của các bản dịch.

Công thức để tính điểm đánh giá của IBM là như sau [4]:

$$score = \exp\left\{\sum_{i=1}^{N} w_i \log(p_i) - \max\left(\frac{L_{ref}}{L_{tra}} - 1{,}0\right)\right\} \qquad (1)$$

- $P_i = \dfrac{\sum_j NR_j}{\sum_j NT_j}$

- NR_j: là số lượng các n-grams trong phân đoạn *j* của bản dịch dùng để tham khảo.
- NT_j: là số lượng các n-grams trng phân đoạn *j* của bản dịch bằng máy.
- $w_i = N^{-1}$
- L_{ref}: là số lượng các từ trong bản dịch tham khảo, độ dài của nó thường là gần bằng độ dài của bản dịch bằng máy.
- L_{tra}: là số lượng các từ trong bản dịch bằng máy.

Giá trị *score* đánh giá mức độ tương ứng giữa hai bản dịch và nó được thực hiện trên từng phân đoạn, ở đây phân đoạn được hiểu là đơn vị tối thiểu trong các bản dịch, thông thường mỗi phân đoạn là một câu hoặc một đoạn. Việc thống kê đồ trùng khớp của các n-grams dựa trên tập hợp các n-grams trên các phân đoạn, trước hết là nó được tính trên từng phân đoạn, sau đó tính lại giá trị này trên tất cả các phân đoạn.

2.2.2 NIST

Phương pháp NIST là sự phát triển trên phương pháp BLEU nhưng có một khác biệt về quan điểm đánh giá là việc chọn lựa n-grams và thông tin trên mỗi n-gram sẽ được sử dụng để phục vụ việc đánh giá.

Sự biến đổi có thể của điểm đánh giá trên một n-gram nếu chúng ta thay đổi vị trí các phần tử trên cùng một n-gram cho chúng ta thấy rằng điểm số cũng sẽ thay đổi nếu chúng ta thay đổi vị trí của các n-grams trên cùng một phân đoạn [2]. Sự thay đổi này sẽ ảnh hưởng lớn lên kết quả đánh giá dựa trên sự tương ứng về vị trí của các n-grams trên phân đoạn. Điều này cho thấy chúng ta có thể sử dụng công cụ số học để tính toán sự biến đổi trên các n-grams bên cạnh sử dụng yếu tố hình học.

Công thức để tính điểm của NIST là [NIST 2002]:

$$score = \sum_{i=1}^{N} \left\{ \sum_{\forall\, w_1 \ldots w_n} \inf(w_1 \ldots w_n) \Big/ \sum_{\forall\, w_1 \ldots w_n \in D_{tra}} \log(p_i) \right\} . \exp\left\{ \beta \log^2 \left[\min\left(\frac{L_{tra}}{L_{ref}}, 1 \right) \right] \right\} \qquad (2)$$

- Những trọng số thông tin là được sử dụng để tính toán trên các n-grams trong tập tất cả các các bản dịch tham khảo theo phương trình sau:

$$\inf(w_1 \ldots w_n) = \log_2 \left(\frac{N_1}{N_2} \right) \qquad (3)$$

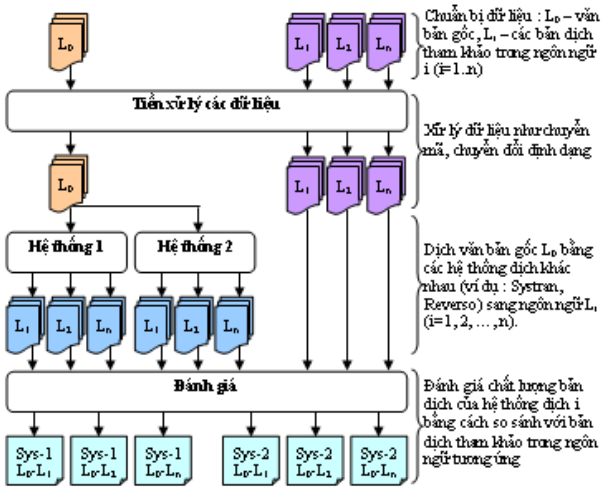
- N_1 = số lượng các tương ứng của các từ $w_1 \ldots w_{n-1}$
- N_2 = số lượng các tương ứng của các từ $w_1 \ldots w_n$
- β là hệ số được chọn bằng 0.5 khi số lượng các từ trong bản dịch máy nhỏ hơn hoặc bằng 2/3 số lượng các từ trong bản dịch tham khảo, ngược lại thì β=1
- N=5
- L_{tra}: số lượng các từ trong bản dịch máy, L_{ref}: số lượng từ trong bản dịch tham khảo.

3 Xây dựng công cụ đánh giá tự động chất lượng hệ thống dịch

Trong phần này chúng tôi giới thiệu công cụ mà chúng tôi phát triển để phục vụ cho việc đánh giá tự động chất lượng các bản dịch.

3.1 Kiến trúc của hệ thống

Để đánh giá chất lượng của một hệ thống dịch trên Internet, chúng ta cần phải gửi một văn bản gốc và một bản dịch tham khảo. Hệ thống của chúng tôi sẽ gửi bản dịch gốc đến các máy chủ phục vụ dịch văn bản ra ngôn ngữ chúng ta cần và lấy kết quả dịch đối chiếu với bản dịch tham khảo để đưa ra kết quả đánh giá về chất lượng bản dịch dựa trên tính điểm đánh giá theo phương pháp BLEU và NIST.



Hình 1. Kiến trúc hệ thống đánh giá chất lượng các hệ thống dịch trên mạng

3.2 Chuẩn bị dữ liệu

Để đánh giá những hệ thống dịch, trước hết chúng ta phải chuẩn bị các dữ liệu đa ngữ (những dữ liệu được viết trong nhiều ngôn ngữ khác nhau như tiếng Anh, Pháp, Nga…) với chất lượng tốt nhất có thể. Chúng ta sẽ sử dụng những dữ liệu này làm văn bản dịch và bản dịch tham khảo (để so sánh với bản dịch máy sau này).

Khi đánh giá, chúng tôi sử dụng 2 nguồn dữ liệu đa ngữ sẵn có đó là: BIBLE - dữ liệu Kinh Thánh (http://bible/gospelcom.net) [8], BTEC – kho dữ liệu đa ngữ hỗ trợ khách du lịch (Basic Travel Expression) [1] để đánh giá các hệ thống dịch. Đây là những dữ liệu đa ngữ có chất lượng rất tốt vì nó được dịch bởi các chuyên gia ngôn ngữ có trình độ cao.
Chúng tôi đã chuẩn bị các dữ liệu như sau:

Bảng 1. Mô tả dữ liệu dùng để đánh giá các hệ thống dịch

Mô tả	BIBLE	BTEC
Số ngôn ngữ	8 (tiếng Anh, Pháp, Trung Quốc, Đức, Tây Ban Nha, Việt Nam, Hy Lạp)	6 (tiếng Anh, Pháp, Trung Quốc, Tây Ban Nha, Đức)
Số lượng câu	32 300 câu / mỗi ngôn ngữ	162 320 câu / ngôn ngữ
Kích thước	5 MB / ngôn ngữ	6 MB / ngôn ngữ

3.3 Xử lý dữ liệu

Trong bước này, chúng tôi đã xử lý những dữ liệu để chuẩn hoá định dạng (chuyển về dạng XML) và mã hoá (chuyển tất cả về mã Unicode). Đây là DTD của các tập tin XML:

```
<!--
 The DTD for evaluation of language translation
-->
<!DOCTYPE MTEVAL [
<!ELEMENT MTEVAL O O (srcset|refset|tstset|DOC+)>
<!ELEMENT (srcset|refset|tstset) - - (DOC+)>
<!ATTLIST (srcset|refset|tstset) setid CDATA #REQUIRED
 srclang (English) #REQUIRED
 trglang (France|Japanese|Chinese|Italia|Arabic) #IMPLIED
>
<!--
 Files of type "srcset" contain source documents to be translated.
 Files of type "refset" contain reference translations to be used in evaluation.
 Files of type "tstset" contain output translations to be evaluated.
-->
<!ELEMENT DOC - - ((hl|p|seg)*)>
<!ATTLIST DOC docid CDATA #REQUIRED
 sysid CDATA #IMPLIED
>
<!ELEMENT hl - - (seg*)>
<!ELEMENT p - - (seg*)>
<!ELEMENT seg - - (#PCDATA)*>
<!ATTLIST seg id CDATA #IMPLIED>
]>
```

3.4 Dịch tự động

Để chiếm được các bản dịch, chúng tôi viết một chương trình để gửi bản gốc đến các máy chủ phục vụ việc dịch và thu hồi lại kết quả. Điều này được thực hiện bởi hàm ***get_doc()*** do chúng tôi viết trong ngôn ngữ lập trình Perl.

Để dịch một văn bản, chương trình của chúng tôi gọi một hàm với các tham số như: URL của máy chủ, nội dung của văn bản cần dịch, tên cặp ngôn ngữ nguồn và đích. Ví dụ, để dịch một văn bản trên Systran, chúng ta gọi ***get_doc()*** như sau:

```
@res=get_doc("www.systranbox.com/systran/box?systran_text=$phrase[$j]&systran_lp=$ls_lc");
```

- **www.systranbox.com/systran/box**: URL của máy chủ Systran
- systran_text = $phrase[$j]: nội dung văn bản cần dịch (1 câu hoặc 1 đoạn)
- systran_lp = $ls_lc: cặp ngôn ngữ nguồn hoặc đích (ví dụ, *$ls_lc = "fr_en"* để dịch từ tiếng Pháp sang tiếng Anh).

4 Thực hiện

Chúng tôi đã phát triển 3 công cụ mà nó phục vụ ở việc đánh giá chất lượng của các hệ thống dịch trên mạng bắt đầu từ những kho dữ liệu có sẵn. Những công cụ này có thể thực hiện trên Internet hoặc trên máy tính đơn (với ngôn ngữ lập trình Perl).

4.1 Chuẩn hoá dữ liệu

Công cụ này cho phép chuyển định dạng một tập tin sang định dạng XML theo một DTD xác định trước. Ví dụ, để chuyển định dạng của kho dữ liệu BTEC từ định dạng TXT (*<Số-hiệu-câu><Nội dung câu>*) sang định dạng XML, chúng tôi viết chương trình và có thể gọi thực hiện như sau:

perl **tranfor.pl** ***–s*** *<tên-tập-tin-nguồn>* ***-t*** *<tên-tập-tin-đích>*

4.2 Dịch tự động

Công cụ này cho phép dịch một văn bản sang một ngôn ngữ được chọn với một hệ thống dịch có sẵn trên Internet. Mục đích của chúng tôi là gọi một hệ thống dịch có sẵn để dịch văn bản được gửi tới qua mạng.

Ví dụ, để dịch một tập tin văn bản tiếng Pháp ***source_fr.txt*** sang tiếng Anh và kết quả sẽ lưu trữ trong tập tin ***tra_en.txt*** bởi hệ thống dịch ***Systran:***

perl **traduire.pl** ***–s*** *source_fr.txt* ***–t*** *tra_en.txt* ***–l*** *fr* ***–c*** *en* ***–m*** *systran*

4.3 Đánh giá các bản dịch

Công cụ này cho phép đánh giá một hệ thống dịch trên mạng và ta cần cung cấp dữ liệu đánh giá dưới 3 tập tin: văn bản nguồn (bản gốc), bản dịch tham khảo (bản dịch tốt nhất có thể của con người) và bản dịch tự động bằng máy.

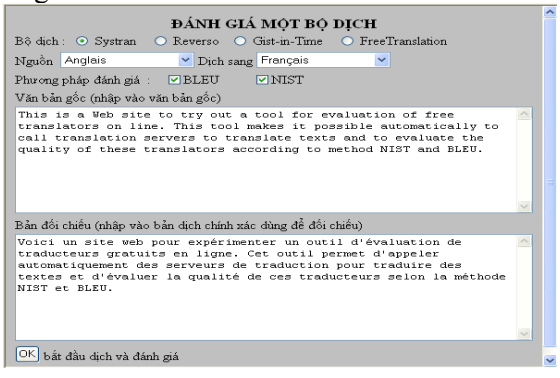
Ví dụ, chúng ta có 3 tập tin: ***src_fr.txt*** (tập văn bản gốc trong tiếng Pháp), ***ref_en.txt*** (tập văn bản tham khảo tiếng Anh, đây là bản dịch tốt của bản gốc tiếng Pháp dùng để đối chiếu) và ***tst_en.txt*** (tập văn bản dịch tự động bởi Systran). Để đanh giá, chúng tôi sử dụng lệnh:

perl **eval.pl** ***–s*** *src_fr.txt* ***–r*** *ref_en.txt* ***–t*** *tst_en.txt*

5 Thử nghiệm

5.1 Công cụ đánh giá

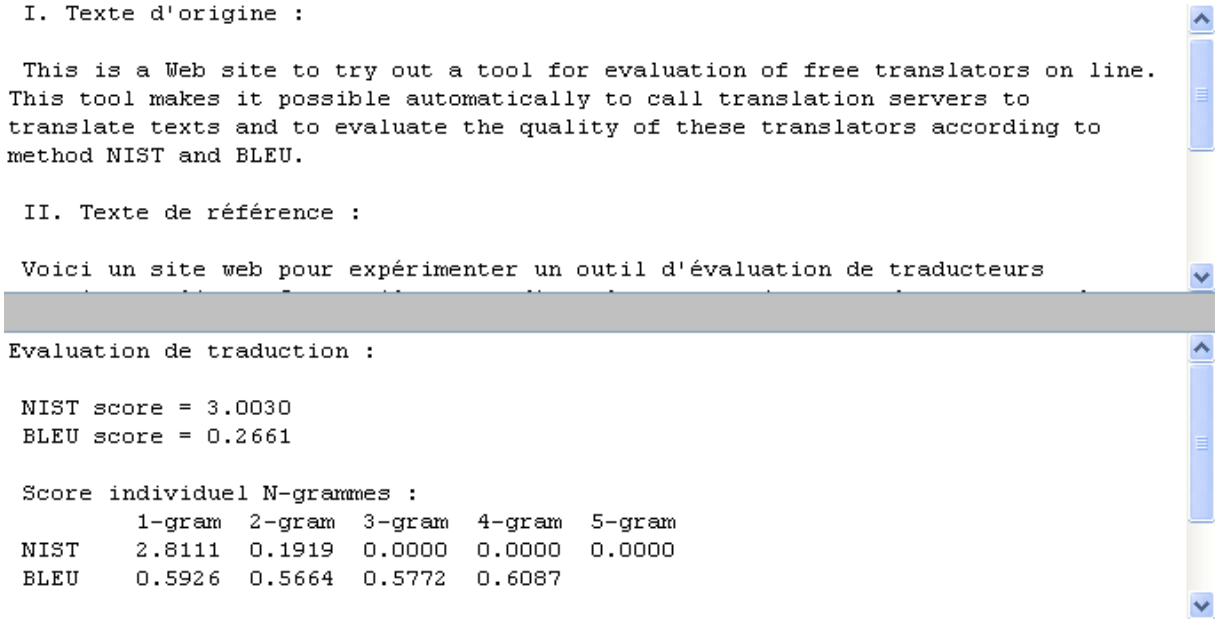
Chúng tôi đã phát triển một trang Web dùng để đánh giá tự động một hệ thống dịch trên mạng. Chúng ta cần phải nhập vào hai văn bản: một văn bản gốc và một bản dịch tham khảo. Tiếp theo, chúng ta chọn một hệ thống dịch cần đánh giá và lựa chọn phương pháp đánh giá theo BLEU hoặc NIST. Giao diện của trang Web như sau:



Hình 2. Giao diện của công cụ đánh giá một hệ thống dịch

Với công cụ này, chúng ta nhận được kết quả đánh giá là độ chính xác của kết quả dịch nhận được từ hệ thống so với bản dịch tham khảo.

Sau khi đánh giá ta có kết quả hiển thị như sau:



Ta lưu ý rằng, với phương pháp NIST nếu điểm số (score) càng lớn (<=10) thì độ chính xác càng cao, ngược lại với BLEU nếu điểm số càng nhỏ (>=0) thì độ chính xác càng cao.

5.2 Kết quả thử nghiệm trên các kho dữ liệu

Chúng tôi đã tiến hành đánh giá chất lượng dịch của các hệ thống dịch miễn phí trên mạng (Systran và Reverso) trên hai kho dữ liệu BTEC và BIBLE.

Chúng tôi giới thiệu một vài kết quả đạt được khi đánh giá với 100000 câu cho mỗi cặp ngôn ngữ được chọn đánh giá:

Bảng 2. Bảng điểm đánh giá Systran và Reverso

Cặp ngôn ngữ	Systran		Reverso	
	BLEU	**NIST**	**BLEU**	**NIST**
Tây Ban Nha→ Anh	0,1322	3,5117	0,1257	3,3567
Anh → Tây Ban Nha	0,0962	3,2985		
Pháp → Anh	0,1277	3,4968	0,1276	3,4010
Anh → Pháp	0,1163	3,1208	0,0996	3,1349

6 Kết luận

Trên cơ sở kết quả đạt được, chúng tôi sẽ tiếp tục thực hiện việc sưu tập dữ liệu và tiến hành đánh giá nhiều hệ thống dịch khác để đưa ra khuyến cáo tốt hơn cho người sử dụng. Tiếp tục nghiên cứu việc dịch trên các hệ thống có sẵn nhưng tham số hóa để tăng độ chính xác, điều này đã thử nghiệm trên các công cụ iTranslator de Bowne Global solutions, IBM và nó cho kết quả dịch tốt hơn.

Trong tương lai, chúng tôi sẽ ứng dụng kết quả này vào một số mục đích như: khuyến cáo người sử dụng nên sử dụng hệ thống dịch nào cho cặp ngôn ngữ cần dịch (nhờ vậy người sử dụng có thể chiếm được kết quả dịch tốt nhất có thể), xây dựng một công cụ trên mạng như là một siêu hệ thống cho phép tự động chọn lựa hệ thống dịch thích hợp cho văn bản của người sử dụng.

TÀI LIỆU THAM KHẢO

[1] BOITET C.: *Approaches to enlarge bilingual corpora of example sentences to more languages*. Papillon-03 seminar, Saporo, July 2003.

[2] Culy C. & Riehemann S.S: *The limits of N-gram translation evaluation metrics*. Proceedings of the Ninth Machine Translation Summit. New Orleans, Louisiana, USA, 2003.

[3] Hovy E.H.: *Toward finely differentiated evaluation metrics for machine translation*. Proceedings of the Eagles Workshop on Standards and Evaluation, Pisa, Italy, 1999.

[4] NIST report: *Automatic evaluation of machine translation quality using N-gram co-occurrene statistics*. http://www.nist.gov/speech/tests/mt/doc/n-gram-study.pdf, 2002.

[5] Popescu-Belis A.: *An experiment in comparative evaluation: Humans vs. Computers*. Proceedings of the Ninth Machine Translation Summit. New Orleans, Louisiana, USA, 2003.

[6] Papineni K., Roukos S., Ward T., Zhu Z-J.: *BLEU: a method for Automatic Evaluation of Machine Translation*. Proceedings of the 20th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, p.p 311-318, July 2001.

[7] Ramaswany G.N., Navratil J., Chaudhari U.V., Zilca R.D: *The IBM system for the NIST 2002 cellullar speaker verification evaluation*. ICASSP-2003, Hong Kong, http://www.reseach.ibm.com/CGB/jiri_pub.html, Avril 2003.

[8] Resnik P., Olsen M.B, Diab M.: *The Bible as a parallel corpus: annotating the "Book off 2000 Tongues"*. Computers and the Humanities, N°33, p.p 129-153, 2000.

[9] Van Slype G.: *Critical study of methods for evaluating the quality of machine translation*. R 19142, http://issco-www.unige.ch/projects/isle/van-slype.pdf, 1979.

[10] VO-TRUNG H.: *Réutilisation de traducteurs gratuits pour développer des systèmes multilingues*, Actes de la conférence RECITAL 2004, Fès, Maroc, Avril 2004.

[11] White J.S, T. O'Connell: *The ARPA MT evaluation methodologies: evolution, lessons, and future approaches*. Proceedings of the first conference of the association for machine translation in the Americas, p.p 193-205, Columbia, Maryland, 1994.