Error Analysis of Named Entity Translation output for Poor-Resourced Bilingual Vietnamese-French Pair

Ngoc Tan Le, Fatiha Sadat

Département Informatique, Université du Québec à Montréal,
201 avenue Président Kennedy, H2X 3Y7 Montréal, Québec, Canada
le.ngoc_tan@courrier.uqam.ca, sadat.fatiha@uqam.ca

Abstract

High-quality translation is time-consuming and an expensive process. Named Entity (NE) Translation, including proper names, remains a very important task for multilingual natural language processing. Most of the gold standard corpora are available for English but not for under-resourced languages such as vietnamese. In Asian languages, this task is remained problematic. This paper focuses on a named entity translation approach by cross-linguistic projection for vietnamese-french, a poor-resourced pair of languages. We incrementally apply a cross-projection method using a small parallel annotated corpora, such as the surface string matching measures according to probabilistic string edit distance similarity and an additional score of syllable consistence feature between the source term and the target term by a syllabification process. Evaluations on vietnamese–french pair show a good accuracy with BLEU gain to 2 points when translating bilingual named entities pairs.

Keywords: Named entity, bilingual corpus, cross-projection, named entity translation, vietnamese-french.

1. Introduction

Due to the multiple meanings of words, expressions and also the metaphors, machine translation systems do not always offer correct translations for given contexts. They may reflect a common name written with upper case as if it is a proper name and vice versa, they translate a name having a signification in a bilingual dictionary as in the case of a common name. Named entity translation, in particular, allows to correctly identify through the proper names languages such as people's names, names of organizations and also the names of the locations. The role of named entity translation is a very important topic in the computational linguistics, statistical machine translation (SMT), cross language information retrieval, information extraction and questions & answers because named entities - particularly persons names, location names and organizations names – inform an essential meaning in natural language processing.

We can see the following examples of correct and incorrect named entity translations from french to vietnamese. Here we use Google Translate to illustrate.
(1.1) [fr] Hier soir, j'ai mangé avec *Monsieur Michel Poulet*. / [en] Last night, I ate with *Mr. Michel Poulet*.
(1.2) [vi] Đêm qua tôi đã ăn *thịt gà với Michel*. [incorrect translation by Google Translate, consulted 11 August 2015]
Literally : Hier soir, j'ai mangé *du poulet avec Michel* (fr) / Last night, I ate *chicken with Michel*. (en)
(1.3) [vi] Tối qua, tôi đã ăn với *ông Michel Poulet*. [correct translation]
In the first example above, the named entity that designates a name of a person was incorrectly translated, "*Monsieur Michel Poulet*" in the sentence (1.1) by "*thịt gà với Michel*" (literally chicken with Michel) in the sentence (1.2) instead of "*với ông Michel*" in the sentence (1.3).
(2.1) [fr] Ma famille voyage dans le delta du *fleuve Rouge*. / [en] My family travels in the delta of the *Red river*.
(2.2) [vi] Gia đình tôi đi ở đồng bằng *sông Đỏ*. [literally incorrect translation]
(2.3) [vi] Gia đình tôi du lịch ở đồng bằng *sông Hồng*. [correct translation by Google Translate, consulted 11 August 2015]
In the second example, the named entity that designates a name of a place has been incorrectly translated "*fleuve Rouge*" in the sentence (2.1) by " *sông Đỏ* " in the sentence (2.2) instead of "*sông Hồng*" in the sentence (2.3). The translation error here is about the synonymy between two words "*Đỏ*" and " *Hồng*" which mean the same signification.
(3.1) [fr] Il est en train de lire *L'Observateur*. / [en] He is reading *L'Observateur*.
(3.2) [vi] Hiện anh đang đọc *The Observer*. [incorrect translation by Google Translate, consulted 11 August 2015]
(3.3) [vi] Anh ta đang đọc *L'Observateur*. [correct translation]
We can see in the third example above, sometimes a discrimination failure between proper name and common name. In this case, the named entity that designates a name of an organization " *L'Observateur* " (3.1) was incorrectly translated from french to vietnamese. His translation was borrowed from english by "*The Observer*" (3.2) instead of keeping it intact name like "*The Observer*" (3.3).

One possible solution is to build a bilingual named entity dictionary. A named entity dictionary or a list of NE pairs is a base for rule-based translation and statistical transliteration method. However, this approach needs firstly a large scale of bilingual corpus with named entity annotated. And manual annotation of bilingual corpora is time-comsuming and an expensive process. Unfortunately, there is very few or no researches regarding the translation of named entities for the couple of french-vietnamese languages.

In this paper we propose an iterative approach to named entity translation by cross-linguistic projection for vietnamese-french, a poor-resourced languages pair.
The structure of this paper is as follows. Section 2 describes the related works about different methods of named entity translation. Section 3 presents our approach about named entity translation for vietnamese-french. Section 4 discusses the experiment setting and results. And conclusion will be given in the last section.

2. Related Works

The task of named entity translation is to translate a named entity, including proper names, temporal and numerical expression from the source language into the target language. Many researchers have tried to solve the named entity translation by several approaches. There are rule-based method, statistical method and web mining method (Ying Liu, 2015).
The rule-based method uses linguistic rules to transliterate and translate named entities. (Stephan Wan et al., 1998) applied this method to transliterate english country names in chinese names.
The statistical method uses a large scale annotated bilingual corpus as training data. And it includes statistical transliteration method, comparable or parallel bilingual corpora-based method. The dominant technique is to create a NE alignment and a bilingual NE lexicon. (Fei Huang and Stephan Vogel, 2002) combined both semantic translation and phonetic transliteration for english-chinese NE translation. (Hassan et al., 2007; Kim et al., 2011; Sellami et al., 2015) proposed, by applying the comparable bilingual corpora-based method, the NE translation based on their context similarity, transliteration similarity and phrase-based translation similarity.
The web mining method uses a large scale of web corpora. (Fei Huang et al., 2005; Long Jiang et al., 2007; Fan Yang et al., 2009; Zhao Mingming et al., 2010) presented a new framework to names translations using web mining method. A given term is submitted to a search engine what extracts the list of translation candidates. This candidate translation list is ranked based on the surface patterns, co-occurrence feature and transliteration feature.
It is challenging to translate named entities across languages with different alphabets and pronunciations such as arabic, russian, korean, japanese, thai, chinese, etc. There are several studies on named entity translation for various language pairs such as english-spanish, french-english, english-arabic, english-japanese, etc. However, we find very few publications on named entity translation for the french-vietnamese, except (Phan, 2014). The machine translation systems face many problems with this pair such as the characteristics of the named entities and the inconsistency of their handwriting and transcription/transliteration in vietnamese. We incrementally apply a cross-projection method using a small parallel annotated corpora, such as the surface string matching measures according to probabilistic string edit distance similarity and an additional score of syllable consistence feature between the source term and the target term by a syllabification process. In this paper, we will present a new framework that deals with the NE translation for french-vietnamese.

3. Our Framework

We present an approach of named entity translation for french-vietnamese. Here we discuss a morphosyntactic appearance between the named entities in the source language and the target language. Our approach relies on two following hypotheses :
Hypothesis 1 : We suppose that a named entity in source language and its translated NE in target language have the same category such as person name, location name or organization name.
Hypothesis 2 : Considering that person and location names are often phonetically translated and their written forms are similar to their pronunciations, we can add an additional syllables consistency feature. It means a syllabification measure comparing the number of syllables in the word blocks or group of words between the source and the target.

Therefore, the proposed approach is composed of three main steps :
Step 1: Extracting a list of french named entities candidates and a list of vietnamese proper names candidates from the french-vietnamese bilingual corpus based on sentences level.
Step 2: Filtering french named entities candidates translated into vietnamese by a statistical model.
Step 3: Scoring a similarity by calculating pairs of bilingual candidates translated by the statistical model. This similarity is based on the Levenshtein string edit distance (1).

$$similarity(S_i, T_j) = 1 - (edit\_distance(S_i, T_j) / maxlength(|S_i|,|T_j|)) \quad (1)$$

The edit distance function of Levenshtein or minimum edit distance is widely used as measurement between two strings. It returns the minimum weight series of edit operations that transforms source word S_i into target word T_j related to the insertion, the deletion or the substitution.

Firstly, the french-vietnamese bilingual corpus is aligned at sentences level. We divide the bilingual corpus into two monolingual corpus as shown in Figure 1. Then, applying a french named entity recognition module, we get a list of French named entity candidates. For vietnamese corpus, we apply a POS (Part-of-speech, grammatical categories) annotation module and we also get a list of vietnamese proper names candidates. Then a statistical model is applied in order to translate the list of french named entity candidates into vietnamese. We obtain a list of translated named entity candidates. And we calculate the similarity scores between this translated named entity candidates as the source and the list of vietnamese proper names candidates as a target. If the value of a pair of scores of bilingual candidates is greater than the threshold value, this pair is chosen and stored in a list of bilingual NE pairs candidates. After analyzing the possible errors of named entities translation, we retrain this post-edited list in the statistical model to observe the possible impacts on the quality of named entities translation.

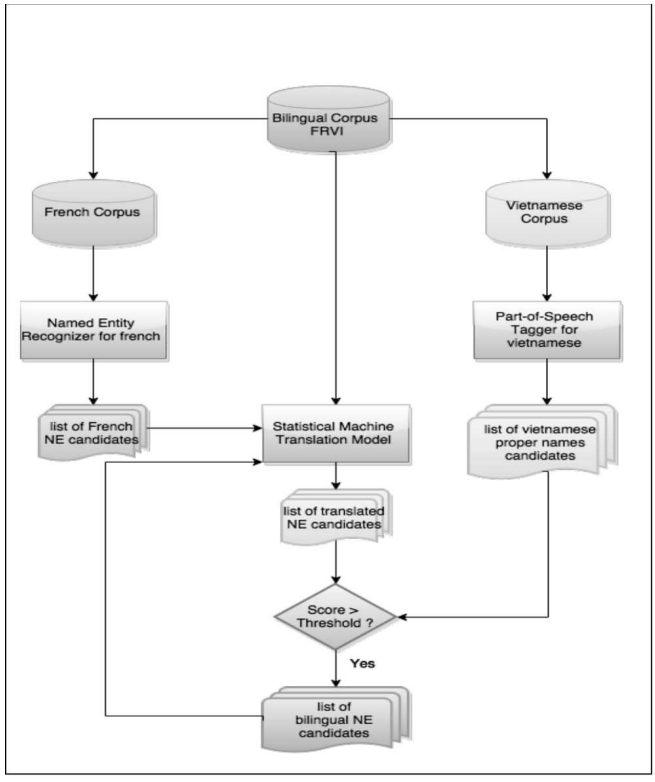

Fig. 1. : Framework of named entity translation system for french-vietnamese

4. Experimentations and Evaluations

The baseline french-vietnamese bilingual corpus is collected with 14,963 sentence pairs from the multilingual web pages for news and 5,284 sentence pairs from the Tam Dao conference, in 2009[1]. This is an international economic conference organized annually in Vietnam, where there are many named entities. Due to time constraints, we only extracted one 2009 version. In perspective we can extract more. We have 1536 named entity including 687 persons names, 713 locations names and 135 organizations names. In our experiments, we use a small test corpus with 1,060 pairs of french-vietnamese pairs phrases. The word segmentation is required for vietnamese corpus. The VCL_WS tool of the VCL group (Vu, 2011) is used for this step. And the vietnamese corpus is also annotated by VCL_POS tagger using the maximum entropy approach (Nguyen, 2011). The automatic annotation system for french is a tool for recognizing named entities developed by (Nouvel et al., 2014).

We implement Moses[2]. This is a statistical machine translation system (Koehn, 2010). Moses offers all the tools needed to build a statistical model. A decoder allows to generate translation assumptions of a source text. It consists of a mechanism capable of effecting the maximization of the equation (2) below in an acceptable time:

$$\text{argmax}_e\ P(e|f) = \text{argmax}_e\ P(f|e)\ P(e) \qquad (2)$$

where :
e : a string in the target language (for example, vietnamese),
f : a string in the source language (for example, French),
p(f|e) : the translation model is the probability that the source string is the translation of the target string,
p(e) : the language model is the probability of seeing that target language string.

To evaluate the named entity translation accuracy, we use the metrics such as BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) and NIST (National Institute of Standards and Technology) (Doddington, 2002) on the test corpus. Table 1 and Figure 2 show the results of the experiments. We performed three following experiments:

• Exp1 (baseline): In this experiment, we use a training data of 14k and a test data of 1,060 bilingual sentence pairs.
• Exp2: In this experiment, we use a training data of 14k and a test data of 359 named entities extracted from 1,060 bilingual sentence pairs.
• Exp3: In this experiment, we use a training data of 14k combining with 5k from Tam Dao 2009 conference and a test data of 359 named entities extracted from 1,060 pairs of bilingual phrases.

	BLEU (%)	**NIST**
Exp1	21.68	5.3977
Exp2	24.05	**3.4556**
Exp3	**26.66**	3.5600

Table 1. : Results of the experimentations of named entity translation for bilingual french-vietnamese

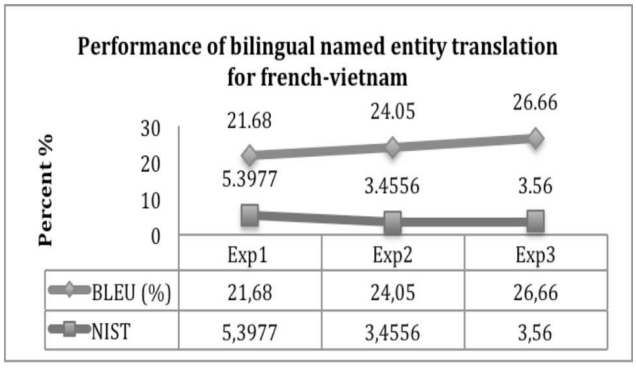

Fig. 2. Performance of bilingual named entity translation for french-vietnamese

Given the above three experiments, we note that the performance is clearly improving in regard to the named entities translation with 21.68% (Exp1 - baseline) -> 24.05% (Exp2, +2.37) -> 26.66% (Exp3, +2.61). We note that there are still unknown words in the first experiment. Hence the metric BLEU was only 21.68% and the metric NIST is 5.3977 due to noise of the unknown words. Then we decide to do the second and third experimentations

[1] http://www.tamdaoconf.com
[2] http://www.statmt.org/moses_steps.html

only with named entities extracted from the test set (including 359 named entities). Thus, the BLEU and NIST metrics show a performance improvement with a gain of more than 2 points of BLEU and NIST over the baseline.

Moreover, we find that the named entities such as person and location names have a great similarity between french and vietnamese. The vietnamese tends to resemble phonetically forms of foreign names with similar pronunciation. We can measure the similarity by counting the number of syllables or the syllables consistency between source term and target term.

For example:

[fr] Phillipines -> [vi] Phi-lip-pin
location name with 3 syllables
[fr] Vietnam -> [vi] Việt Nam
location name with 2 syllables
[fr] Singapour -> [vi] Sing-ga-po
location name with 3 syllables
[fr] Pharaon -> [vi] Pha-ra-ông
person name with 3 syllables
[fr] Joseph -> [vi] Giô-sép
person name with 2 syllables

In addition, we find that there are shortcomings and errors in the bilingual named entity translation. A major drawback of a system based on the statistical machine translation model involves the amount of training data. The training data should be as large as possible in order to cover all linguistic varieties of translations.

Translation errors are categorized into three criteria : lexicon, syntax and transcription / transliteration.

1) Lexical errors concern the lack of words. The system deals with the out-of-vocabulary words, the missing words or the incorrect words. For example:

[fr] fleuve Rouge -> [vi] sông Đỏ
instead of « sông Hồng »
[fr] Mékong -> [vi] Mékong
instead of « Cửu Long »
[fr] Long Biên -> [vi] Dài Biên
instead of « Long Biên »

2) Syntax errors concern the mistranslation of names structures or word order in noun phrases. For example:

[fr] Asie du Sud-Est -> [vi] Á của Đông Nam
instead of « Đông_Nam_Á »
[fr] Afrique -> [vi] Phi Châu
instead of « Châu Phi »
[fr] Asie orientale -> [vi] Châu Đông_Á
instead of « Á Đông »

3) Transcription / transliteration errors relate to proper names which are poorly transcribed or transliterated by the machine translation system due to the influence of english words during manual translation. For example:

[fr] Singapore -> [vi] Singapore
instead of « Singapour -> Sing-ga-po »
[fr] Algeria -> [vi] An-giê-ri
instead of « Algérie »
[fr] Californie -> [vi] California
instead of « Ca-li-pho-ni-a »
[fr] Malaysie -> [vi] Malaysia
instead of « Malaisie -> Ma-lai-xi-a » or « Mã Lai »

5. Conclusion and Perspective

In this paper we presented an approach on named entity translation by cross-linguistic projection for vietnamese-french, a poor-resourced pair of languages. We applied a cross-projection method using a small parallel annotated corpora, and calculating the surface string matching measures according to probabilistic string edit distance similarity and an additional score of syllable consistence feature between the source term and the target term by a syllabification process. Evaluations on vietnamese–french pair of languages show a good accuracy with BLEU gain to 2 points when translating bilingual named entities pairs. This resulted in a small bilingual annotated corpus in a significant improvement into named entity translation.

In perspective, we will focus on collecting a large scale bilingual corpus. We will deal with different kinds of error in NE translation and propose to introduce other features (i.e. features based on transliteration) in order to improve the quality of the extracted NE translation. The framework can be naturally extended to comparable corpora of more than two languages.

References

Ying Liu (2015). *The technical analyses of Named Entity Translation*. International Symposium on Computers & Informatics, ISCI 2015, pp. 2028-2037.

Stephan Wan, Cornelia Maria Verspoor (1998). *Automatic english-chinese name transliteration for development of multilingual resources*. Proceedings of the 17th international conference on Computational linguistics, volume 2, pp. 1352-1356.

Fei Huang, Stephan Vogel (2002). *Improved Named Entity Translation and Bilingual Named Entity Extraction*. Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces, ICMI'2002.

Hassan, A. Fahmy, H. and Hassan, H.(2007). *Improving Named Entity Translation by Exploiting Comparable and Parallel Corpora*. Proceedings of the 2007 Conference on Recent Advances in Natural Language Processing (RANLP), AMML Workshop.

Kim, J. Jiang ; L. Hwang, S. Song, Y. and Zhou, M.(2011). *Mining Entity Translations from Comparable Corpora: A Holistic Graph Mapping Approach*. Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM 2011), pp.1295-1304.

Rahma Sellami, Fatiha Sadat and Lamia Belguith Hadrich (2015). *Mining Named Entity Translation from Non Parallel Corpora*. Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference, pp.219-224.

Fei Huang, Ying Zhang and Stephan Vogel (2005). Mining key phrase translation from web corpora, Proceedings of the HLT/EMNLP-2005, pp. 483-490.

Long Jiang, Ming Zhou, Lee-Feng Chien and Cheng Niu (2007). *Named Entity Translation with web mining and transliteration*. IJCAI-07, pp. 1629-1634.

Fan Yang, Jun Zhao and Kang Liu (2009). *A Chinese-English Organization Name Translation System Using Heuristic Web Mining and Asymetric Alignment*.

Proceedings of the 47th Annual meeting of the ACL and the 4th IJCNLP of the AFNLP, pp. 387-395.

Zhao Mingming, Hong Yu and You Jianmin (2010). *Research on Name Entity Translation Based on Transliteration and Web*. Proceedings of the 6th national conference on Information Retrieval, pp. 357-366.

Phan Thi Thanh Thao (2014). *Machine translation of proper names from english and french into vietnamien : an error analysis and some proposed solutions*, Ph.D. thesis, 11 March 2014.

Vu Dinh Hong (2011). *Phân đoạn từ tiếng việt ngữ dụng*. Master Thesis of University of Sciences, National University of Ho Chi Minh city.

Nguyen Khuong An, Dinh Dien (2011), *Tích hợp thông tin từ loại vào hệ dịch máy thống kê*. National Conference, Cần Thơ, 2011.

Damien Nouvel, Jean-Yves Antoine and Nathalie Friburger (2014). *Pattern Mining for Named Entity Recognition*. LNCS/LNAI Series volume 8387i (post-proceedings LTC 2011) 2014.

Philipp Koehn (2010). *Statistical Machine Translation*. Cambridge University Press, New York, ISBN-13 978-0-521-87415-1 Hardback 2010.

Papineni, K., Roukos, S., Ward, T., Zhu, W. J. (2002). *BLEU: a method for automatic evaluation of machine translation*. ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311–318. CiteSeerX: 10.1.1.19.9416.

Doddington, G. (2002). *Automatic evaluation of machine translation quality using n-gram cooccurrence statistics*. Proceedings of the Human Language Technology Conference (HLT), San Diego, CA pp. 128–132