Building a Bilingual Vietnamese-French Named Entity Annotated Corpus through Cross-Linguistic Projection

Ngoc Tan Le[1], Fatiha Sadat[1]
(1) 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

Résumé. La création de ressources linguistiques de bonne qualité annotées en entités nommées est très coûteuse en temps et en main d'œuvre. La plupart des corpus standards sont disponibles pour l'anglais mais pas pour les langues peu dotées, comme le vietnamien. Pour les langues asiatiques, cette tâche reste très difficile. Le présent article concerne la création automatique de corpus annotés en entités nommées pour le vietnamien-français, une paire de langues peu dotée. L'application d'une méthode basée sur la projection cross-lingue en utilisant des corpus parallèles. Les évaluations ont montré une bonne performance (F-score de 94.90%) lors de la reconnaissance des paires d'entités nommées dans les corpus parallèles et ainsi la construction d'un corpus bilingue annoté en entités nommées.

Abstract. The creation of high-quality named entity annotated resources is time-consuming and an expensive process. Most of the gold standard corpora are available for English but not for less-resourced languages such as Vietnamese. In Asian languages, this task is remained problematic. This paper focuses on an automatic construction of named entity annotated corpora for Vietnamese-French, a less-resourced pair of languages. We incrementally apply different cross-projection methods using parallel corpora, such as perfect string matching and edit distance similarity. Evaluations on Vietnamese –French pair of languages show a good accuracy (F-score of 94.90%) when identifying named entities pairs and building a named entity annotated parallel corpus.

Mots-clés : Entité nommée, corpus parallèle, projection cross-lingue.
Keywords: Named entity, parallel corpus, cross-projection.

This demonstration concerns a Named Entity Recognizer (NER) tool for Vietnamese-French, a less-resourced pair of languages using bilingual parallel corpora. Our approach is based on the assumption that a word A (or a phrase A') in the source language L_1 is often translated to a word B (or a phrase B') in the target language L_2. Thus, we can expect that their translations also co-occur more often in the target language (Fung, 2000). Based on this assumption, different steps were proposed for an automatic extraction of bilingual NER pairs in a parallel corpus and the construction of a NER tool for Vietnamese-French, a less-resourced pair of languages.

Our proposed strategy of best matching criterion for the extraction of bilingual NER pairs from parallel corpora relies on the following steps. First, for each word w_i in the source and target corpora, we build context vectors by considering a window size in the two corpora, the content words, the label and the co-occurrence of all words. The context vectors of the target words are translated using a bilingual dictionary and some gazetteers. Finally, a similarity for each source term S and for each target term T is computed on the basis of the Leveinshtein distance with S_1 and T_1 are a word or a phrase respectively in source term S and in target term T:

$$similarity(S_1,T_1) = 1 - \frac{edit\_distance\,(S_1,T_1)}{maxlength(|S_1|,|T_1|)} \qquad (1)$$

As a preprocessing step, we tag and lemmatize the text in both languages. The bilingual French-Vietnamese corpora contain 20,000 pairs of sentences. This step allows us to focus on content words only (nouns, verbs, adjectives and adverbs) and thus reduces the noise in our model. Content words are the primary focus for thesaurus enrichment. Word segmentation for Vietnamese is completed using the VCL_WS tool of the VCL group (Vu et al., 2011). Moreover, the Vietnamese corpus is annotated using the VCL_POS tagger, which relies on the maximum entropy approach (Nguyen et al., 2011). The TagEN tool (Taggueur Entités Nommées – Named Entity Tagger) for French is a tool for recognizing named entities developed by Jean-François Berroyer and Thierry Poibeau at Laboratoire d'Informatique de Paris-Nord (LIPN) (Poibeau, 2003).

In this demonstration, our interest concerns the automatic extraction of NE pairs using a cross-linguistic method and thus the construction of bilingual NE lexicons of proper names for location, organization and person. We evaluate the quality of the bilingual NE pairs for French-Vietnamese bilingual sentence pairs with the help of linguistic experts.

Our evaluation uses a test file of 1,060 pairs of French-Vietnamese sentences pairs and shows a good accuracy (F-score of 94.90%). The bilingual extracted NE pairs can be used to enrich a bilingual dictionary and gazetteers. Table 1, Figures 1 and Figure 2 show the results using the precision, recall and F-score values of the developed NER tool.

The developed NER tool will be used in a complete information extraction system, which can be used in many NLP and machine learning applications such as text classification, machine translation, information retrieval, summarization, etc.

	#Correct-Found	**#Noise**	**#To find**	**Precision**	**Recall**	**F-score**
Location	237	4	246	98.34%	96.34%	97.33%
Organization	17	11	55	60.71%	30.91%	40.96%
Person	86	4	96	95.56%	89.58%	92.47%
			Average	**96.95%**	**92.96%**	**94.90%**

TABLE 1 : Results of the experimentation with a test set of 1,060 bilingual French-Vietnamese sentences pairs

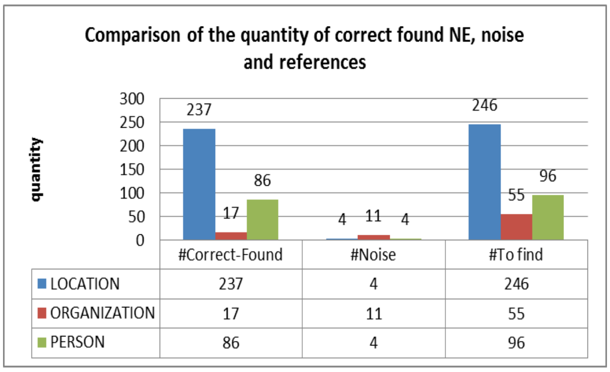

FIGURE 1: Comparison of the quantity of correct found NE, noise and references

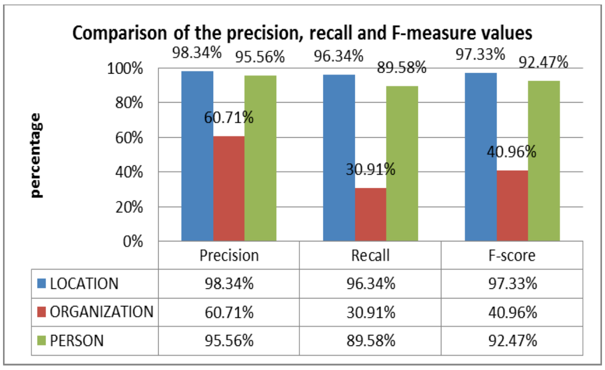

FIGURE 2: Comparison of the precision, recall and F-measure

References

CHINCHOR, N. (1998). MUC-7 named entity task definition. *In Proceedings of the 7th Message Understanding Conference*.

FUNG, P. (2000). A statistical view of bilingual lexicon extraction: From parallel corpora to non-parallel corpora. *In Jean Véronis, editor, parallel text processing*.

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*.

THIERRY POIBEAU (2003). The multilingua named entity recognition framework. *Association for Computational Linguistics*, 155-158.