How a low-resource named entities recognition and transliteration framework for Vietnamese can improve the automatic machine translation ?

Ngoc Tan Le[1] · Fatiha Sadat[1]

[1]Department of Computer Science, Universite du Quebec a Montreal, Montreal, Canada
le.ngoc_tan@courrier.uqam.ca · sadat.fatiha@uqam.ca (https://github.com/NgocTanLE)

Motivation

- **Named Entity Recognition**: a subtask of information extraction that seeks to locate and classify named entity mentions in unstructured text into pre-defined categories such as the person names, organizations, locations, etc.
- **Transliteration**: the process of converting a text in one script to another, guided by phonetic clues (Knight and Graehl, 1998)
- Transliteration considered as a sub-task of machine translation (MT)
- **Research Objectives**:
  - ☐ Deal with out-of-vocabulary words (OOV), considered as proper names or technical terms, derived from MT system

Proposed Approach

In the statistical approach, our models apply the phrase-based architecture. In the deep learning approach, our models apply the encoder-decoder recurrent neural networks (RNN) architecture, with Long-Short Term Memory (LSTM) (Hochreiter andSchmidhuber, 1997).

Named Entity Recognition System

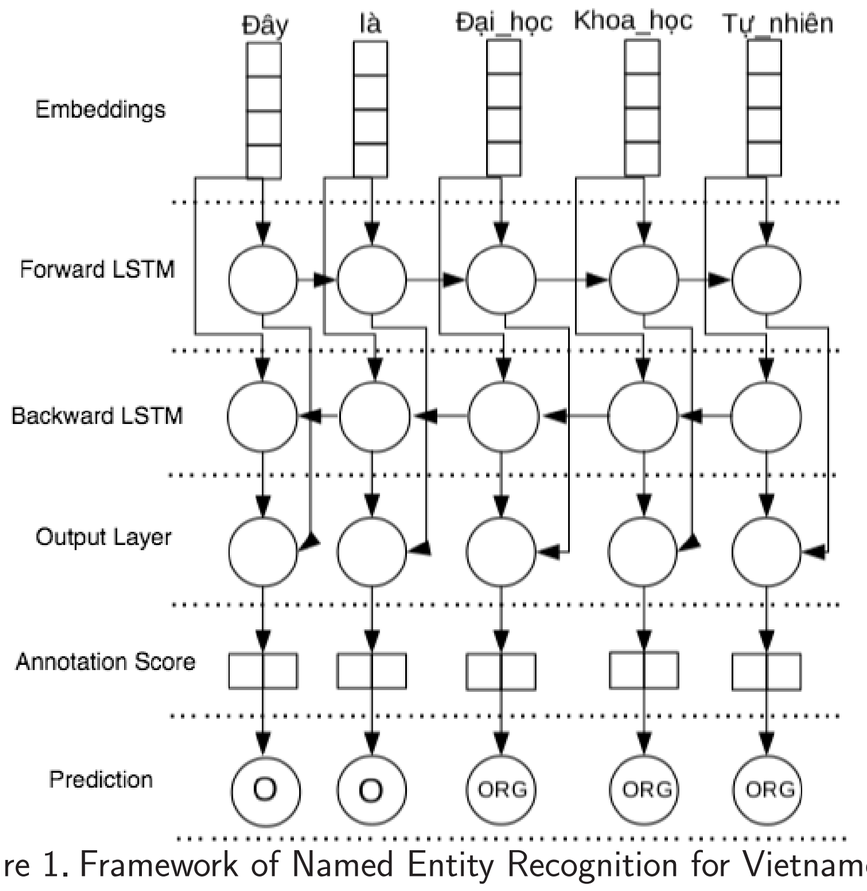

Figure 1. Framework of Named Entity Recognition for Vietnamese

Machine Transliteration

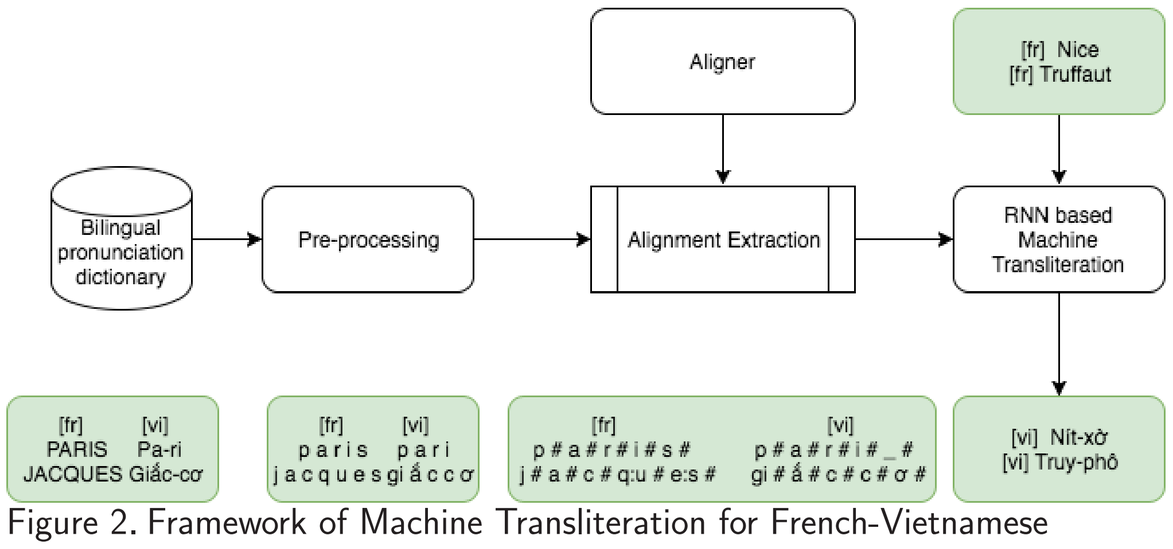

Figure 2. Framework of Machine Transliteration for French-Vietnamese bilingual named entities

Machine Translation

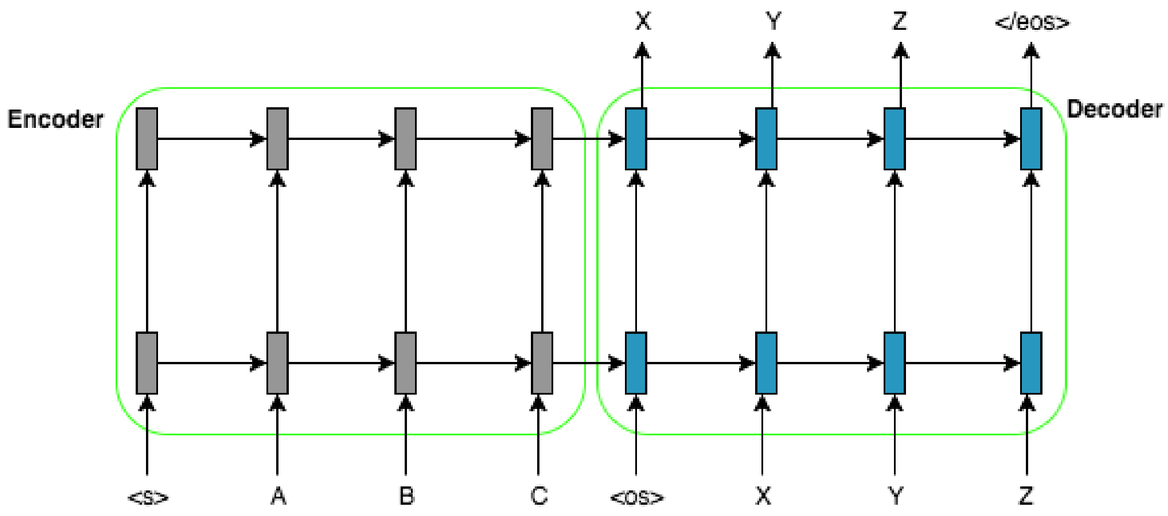

Figure 3. An encoder-decoder RNN-based model architecture with two layers, illustration depicted from (Sutskeveret al., 2014).

Experiments and Results

NER System

Experiments	P (%)	R (%)	F1 (%)
System 1 (SVM)	85,23	78,02	81,46
System 2 (CRF)	**86,70**	79,54	82,97
System 3 (Bi-LSTM, w/o features)	81,08	83,50	82,27
System 4 (our approach)	84,53	**87,93**	**86,20**

Table 1. Performances of NER systems

Machine Transliteration

Experiments	BLEU ↑	TER ↓	GER ↓
Baseline (pbSMT)	61.30	24.08	44.20
System 1 (enc-dec)	92.38	9.69	18.28
System 2 (enc-dec)	94.41	4.70	8.87
System 3 (enc-dec)	**95.96**	**3.28**	**6.19**

Table 2. Evaluation of Machine Transliteration systems

Machine Translation

Experiments	BLEU	METEOR	TER	OOV
System 1 (pbSMT)	31.40	40.50	67.6	49.80
System 2 (pbSMT)	40.00	63.40	49.70	38.60
System 3 (pbSMT)	**51.80**	**68.40**	**36.20**	**27.50**
System 4 (NMT)	12.04	28.30	91.94	69.72

Table 3. Evaluations of Machine Translation systems of named entities for French-Vietnamese

Conclusion and Perspectives

- Promising results : The *RNN-based* system outperformed both the *phrasal SMT* in NER and Machine Transliteration.
- Perspectives :
  - ☐ Experiment with a larger bilingual phonetic corpus