Detecting Non-literal Translations by Fine-tuning Cross-lingual Pre-trained Language Models

†Yuming Zhai, ♣Gabriel Illouz, ♣Anne Vilnat

†BFSU Artificial Intelligence and Human Languages Lab, Beijing Foreign Studies University, 100089, Beijing, China

♣ University of Paris-Saclay, CNRS, LIMSI, 91400, Orsay, France

Objectives

- Automatically detecting non-literal translations at sentence and phrase level.
- Help to construct materials to teach translation to human learners.
- Serve as the first step to assembling as much representative data as we can to train MT systems to start producing more non-literal translations than their literal alternatives.

Introduction

- Non-literal translations could bring difficulties for NLP and inappropriate ones could be undesirable for certain tasks.
- Nonethelss, non-literal but appropriate translations are difficult to produce and machines are still on the way to simulate human translators on this aspect.
- In order to foster the study on non-literal translations, automatically detecting them in parallel corpora is an important step.
- Our approach is based on the advances in cross-lingual pre-trained language models: XLM (Conneau and Lample, 2019).

Research Questions

- Do non-literal translations occur more often in human translations than in machine translations?
- Could pre-trained language models be fine-tuned to detect the presence of non-literal translations at sentence level?
- Could the architecture be adapted to distinguish literal and non-literal translations at phrase level?

Dataset

- Different types of non-literal translations are formalized by translation techniques in translation studies (Vinay and Darbelnet, 1958; Chuquet and Paillard, 1989). We use the English-French corpus of TED Talks annotated with translation techniques at sub-sentential level (Zhai et al., 2019).

At sentence level

- We first trained a Human vs Machine translation classifier, by fine-tuning XLM with four corpora.
- The machine-translated sentences are generated by using Fairseq (Ott et al., 2019).
- We show that there exists a moderately positive correlation between the prediction probability of human translation and the non-literal translations' proportion in a sentence.
- Directly fine-tuning XLM on this dataset obtains a better accuracy than the majority vote baseline.
- Resuming the fine-tuning after loading the final trained Human-vs-Machine translation classifier model on Literary books and Europarl corpus provides a gain of performance.

Majority vote baseline	72.91%
Approach	**Average best validation accuracy**
Directly fine-tune XLM	78.66% ± 3.93%
Resume fine-tuning after loading the final trained human-vs-machine translation classifier model:	
Literary books	**80.16%** ± 3.96%
Europarl	<u>79.86%</u> ± 3.45%
OpenSubtitles	78.07% ± 3.11%
Ted Talks	78.24% ± 3.83%

Figure 1: Detecting the presence of non-literal translations in a sentence. The best performance is in bold, the second best is underlined

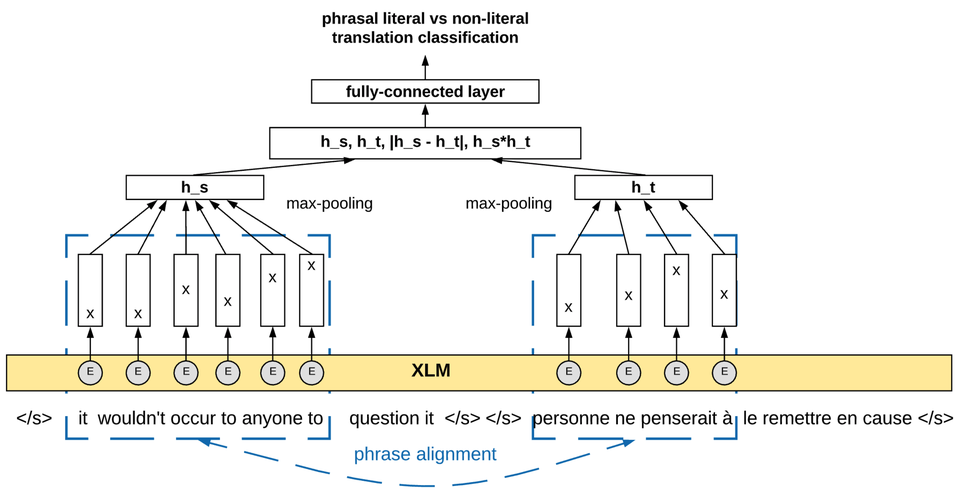

Figure 2: Fine-tune XLM at phrase level to classify literal vs non-literal translation

At phrase level

- This step is inspired by the work of Arase and Tsujii (2019).
- The architecture is presented in Figure 2. Preprocessing steps: lowercasing, accent removing, BPE sub-word tokenization.
- Our dataset provides the token indexes of aligned phrases for each sentence pair. The original alignment indexes of each phrase pair are adjusted after all the preprocessing steps.
- We compare the performance of this XLM-based classifier with a RandomForest classifier which leverages 198 linguistic features (Zhai et al.,2019). The test accuracy is 85.20% and 90.90%, respectively.
- The oracle study shows that there exists a complementarity between the two methods, therefore a hybrid classifier could be investigated in future work to improve the performance.

Conclusion and future work

- Generic sentence representations produced by XLM are transferable to our task after fine-tuning.
- Leverage the advances of automatic word alignment to reduce the reliance on manual work at phrase level.

Contact Information

- Personal page: http://yumingzhai.github.io/
- Email: zhaiyuming@bfsu.edu.cn