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

Yuming Zhai
BFSU Artificial Intelligence and Human Languages Lab
Beijing Foreign Studies University
100089, Beijing, China
zhaiyuming@bfsu.edu.cn

Gabriel Illouz
Université Paris-Saclay
CNRS, LIMSI
91400, Orsay, France
illouz@limsi.fr

Anne Vilnat
Université Paris-Saclay
CNRS, LIMSI
91400, Orsay, France
vilnat@limsi.fr

Abstract

Human-generated non-literal translations reflect the richness of human languages and are sometimes indispensable to ensure adequacy and fluency. Non-literal translations are difficult to produce even for human translators, especially for foreign language learners, and machine translations are still on the way to simulate human ones on this aspect. In order to foster the study on appropriate and creative non-literal translations, automatically detecting them in parallel corpora is an important step, which can benefit downstream NLP tasks or help to construct materials to teach translation. This article demonstrates that generic sentence representations produced by a pre-trained cross-lingual language model could be fine-tuned to solve this task. We show that there exists a moderate positive correlation between the prediction probability of being human translation and the non-literal translations' proportion in a sentence. The fine-tuning experiments show an accuracy of 80.16% when predicting the presence of non-literal translations in a sentence and an accuracy of 85.20% when distinguishing literal and non-literal translations at phrase level. We further conduct a linguistic error analysis and propose directions for future work.

1 Introduction

Translation is a cross-lingual and intercultural process helping to communicate across language barriers. Human-generated parallel corpora have been largely exploited to develop machine translation (MT) systems, which rapidly evolve and are widely used in production nowadays (Koehn et al., 2003; Och and Ney, 2003; Wu et al., 2016; Gehring et al., 2017; Vaswani et al., 2017; Hassan et al., 2018). Multiple human-generated references could be used to evaluate each MT hypothesis, since the same meaning could be expressed in different ways.

When we take a closer look at human translations, apart from literal translations, fixed corresponding expressions or the most commonly used renderings, various translation techniques could be employed to produce non-literal translations. The resulting variations reflect the richness of human languages. Furthermore, because of the existing differences between languages and cultures, it is sometimes inevitable to translate non-literally to ensure adequacy and fluency. Table 1 presents two English-French translations found in the subtitles of TED Talks[1], produced by volunteer translators. We mark the non-literal translations in bold and present the literal translation of the French rendering in brackets. As shown below, non-literal translations can occur at word, phrase and even sentence level.

Different techniques of producing non-literal translations are systematically studied by linguists and translation scholars, such as *generalization, particularization, modulation, transposition,* etc. (Vinay and Darbelnet, 1958; Chuquet and Paillard, 1989; Molina and Hurtado Albir, 2002). Based on these studies, Zhai (2019) investigated the possibility of annotating and automatically recognizing different translation techniques at sub-sentential level. For natural language processing (NLP) tasks, non-literal translations can bring difficulties for automatic word alignment (Dorr et al., 2002; Deng and Xue, 2017) or for paraphrase extraction via bilingual pivoting (Bannard and Callison-Burch, 2005; Pavlick et al.,

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

[1] https://www.ted.com/

Proceedings of the 28th International Conference on Computational Linguistics, pages 5944–5956
Barcelona, Spain (Online), December 8-13, 2020

It's nothing if not ambitious. → ***C'est véritablement ambitieux.*** (Lit. It's really ambitious.)
That's why ***we've made it our primary goal*** *to* ***spend a large sum of money*** *on an* ***advertising effort to help bring out*** *and* ***complicate*** *the truth about coal.* → *C'est pourquoi* ***nous avons décidé*** *de* ***dépenser massivement*** *pour une* ***campagne de communication destinée à dire****, et à* ***brouiller****, la vérité sur le charbon.* (Lit. That's why we decided to spend massively on a communication campaign intended to tell, and blur, the truth about coal.)

Table 1: English-French non-literal translations, found in the subtitles of TED Talks

2015). Non-literal translations can also cause noisy sentence pairs in parallel corpora, which affect the training of MT systems (Carpuat et al., 2017; Pham et al., 2018; Vyas et al., 2018). On the other hand, non-literal but appropriate translations are difficult to produce (Carl and Schaeffer, 2017) and machines are still on the way to simulate human translators on this aspect (Ahrenberg, 2017; Toral and Way, 2018). To inspire MT system's development, efforts have been done to analyze language contrasts through alignment discrepancies (Lapshinova-Koltunski and Hardmeier, 2017), and to detect free and fluent translation examples from English-Chinese parallel corpora (Chen et al., 2018).

In order to foster the study on non-literal translations, automatically detecting them in parallel corpora is an important step, which can help constructing materials to teach translation to human learners or 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. Our research questions are the following: do non-literal translations occur more often in human translations? Could pre-trained language models be fine-tuned to detect the presence of non-literal translations in a sentence? And could the architecture be adapted to distinguish literal and non-literal translations at phrase level?

To answer these questions, our approach is based on the advances in cross-lingual pre-trained language models. Recently, Conneau and Lample (2019) introduced a new supervised learning objective that improves cross-lingual language model's pre-training when parallel data is available.[2] Having a neural architecture based on Transformer (Vaswani et al., 2017), their pre-trained cross-lingual LMs (XLMs) provide general-purpose text representations, which can encode any sentence into a shared embedding space. For our task of detecting non-literal translations, we use an English-French corpus of TED Talks annotated with translation techniques at sub-sentential level (Zhai et al., 2019), and we demonstrate that the generic sentence representations produced by XLM are transferable to our task after fine-tuning.

2 Related work

Research at the intersection of translation studies and NLP around non-literal translations has attracted the attention of many researchers. In translation theories, different typologies of translation techniques are proposed to formalize human translators' choices when they translate non-literally (Vinay and Darbelnet, 1958; Chuquet and Paillard, 1989; Molina and Hurtado Albir, 2002). With established annotation guidelines and an adapted typology, Zhai (2019) annotated EN-FR and EN-ZH parallel corpora with translation techniques, and conducted multi-class classification of translation techniques.

Non-literal translations could bring difficulties for NLP and inappropriate ones could be undesirable for certain tasks. Deng and Xue (2017) semi-automatically identified, categorized and quantified seven types of translation divergences between Chinese and English, which arise either because of cross-lingual differences or because of non-literal translations. Various models are proposed to automatically detect translation divergences in parallel corpora, with the goal of filtering out divergent sentence pairs to improve MT systems' performance (Carpuat et al., 2017; Pham et al., 2018; Vyas et al., 2018). In the process of bilingual pivoting paraphrasing (Bannard and Callison-Burch, 2005), non-literal translations could be a reason leading to diverse semantic relations in the resource PPDB 2.0 apart from true paraphrases (Pavlick et al., 2015).

[2] https://github.com/facebookresearch/XLM

On the other hand, non-literal but appropriate translations are not negligible and are studied by different approaches. Carl and Schaeffer (2017) investigated the effects of cross-lingual syntactic and semantic distance on translation production times and found that non-literality makes from-scratch translation and post-editing difficult. In a case study of comparing human and machine translations, Ahrenberg (2017) suggested that the human translator used several procedures that seem to be beyond the reach of the MT system (such as sentence splitting, shifts of phrase function and/or category, explicitation, modulation and paraphrasing). The project of Fraisse et al. (2019) aims to preserve cultural heritage and language diversity by analyzing the translation adaptations in multilingual corpora of translated literary texts, which is particularly important for low-resource languages. In order to provide insights for discourse-aware MT system's development, discourse-related language contrasts are analyzed for English-Croatian and English-German (Lapshinova-Koltunski and Hardmeier, 2017; Šoštarić et al., 2018). For inspiring MT's further improvement on fluency and for human translators' reference, Chen et al. (2018) proposed a method for detecting free translation examples from bilingual parallel corpora, which is based on an innovative use of attention scores. Yuan and Sharoff (2020) proposed a stacked neural network for fine-grained human translation quality estimation, and they discussed that this model has limited validity for adequate scoring of free but still valid translations.

In this paper, by using the English-French TED Talks corpus annotated with translation techniques by Zhai et al. (2019), we investigate whether pre-trained cross-lingual language models could be fine-tuned to detect non-literal translations at sentence and phrase level. For the latter, we compare the results with those obtained by Zhai et al. (2019).

3 Detecting non-literal translations at sentence level

Our first goal is to detect whether there are non-literal translations in a sentence. By assuming that human translators employ more non-literal translations than machines (Toral and Way, 2018), we first transfer the problem into training a model to distinguish human and machine translations (Rarrick et al., 2011), expecting that the classifier would learn the linguistic differences between them and further help predict the presence of non-literal translations in a sentence.

After training this human-vs-machine translation classifier, we investigate whether there is a positive correlation between the prediction probability of human translation and the non-literal translations' proportion in a sentence. Finally, we test the hypothesis that resuming the fine-tuning task on detecting the presence of non-literal translations in a sentence after loading this human-vs-machine translation classification model could get better performance.

3.1 Human or machine translation classifier

XLM pre-trained model Conneau and Lample (2019) pre-trained cross-lingual language models with three objectives: CLM (Causal Language Modeling), MLM (Masked Language Modeling)[3], and MLM in combination with TLM (Translation Language Modeling, an extension of MLM, i.e. randomly masking words in both the source and target sentences)[4]. For TLM, parallel sentences are concatenated as input. When predicting a masked word in an English sentence, the model can either attend to surrounding English words or to the French translation, encouraging the model to align EN an FR representations.

In our experiments, we fine-tune XLM's released model *mlm_tlm_xnli15_1024.pth*, which is pre-trained with the objectives MLM+TLM. Conneau and Lample (2019) used 80k BPE (Byte Pair Encoding (Sennrich et al., 2016)) splits and a vocabulary of 95k sub-word units, and trained a 12-layer bidirectional Transformer model (1024 hidden states) on the Wikipedias of 15 languages of the XNLI dataset (Conneau et al., 2018). This model can be used to obtain a better initialization of sentence encoders for zero-shot cross-lingual classification, as is demonstrated in their fine-tuning experiment on XNLI benchmark (Cross-lingual Natural Language Inference).

Our fine-tuning scheme In our case, the fine-tuning is conducted on a dataset of EN-FR Human

[3]Differences between their approach and the MLM of BERT (Devlin et al., 2019) include the use of text streams of an arbitrary number of sentences (truncated at 256 tokens) instead of pairs of sentences.

[4]This produces a supervised cross-lingual LM that combines both the MLM and the TLM loss using additional parallel data.

vs Machine translations. Since the classifier will be later applied on a sub-corpus of TED Talks, we choose TED Talks, OpenSubtitles, Literary Books[5] (Tiedemann, 2012) and Europarl (Koehn, 2005) to observe the effects of having similar and different training corpus genre. The statistics of the four corpus are shown in Table 2. The original French text is produced by human translators, and the machine-translated sentences are generated by using fairseq (Ott et al., 2019)[6] with their pre-trained model *transformer.wmt14.en-fr* (Ott et al., 2018).[7] The de-tokenized BLEU scores calculated by SacreBLEU (Post, 2018) and CHRF3 (Popović, 2015) for each machine-translated corpus are in Table 3.

	Nb sentence pairs	Nb English tokens
Literary books	33 669	876 866
Europarl	30 000	869 869
OpenSubtitles	30 000	215 584
TED Talks	30 000	498 440

Table 2: Statistics of different corpora used to train human-vs-machine translation classifier. 30k sentence pairs are randomly taken for the last three corpora

Conneau and Lample (2019) processed all languages with the same shared vocabulary created through BPE, which greatly improved the alignment of embedding spaces across languages. Therefore, after tokenizing input sentences with the Moses tokenizer (Koehn et al., 2007), we use fastBPE[8] to split tokens into sub-word units with the pre-learned BPE splits. The processed sentence pairs (source sentence concatenated with human or machine translation, in form of tensors) are fed to the XLM model to generate sentence embeddings (the sentence length is clipped to have a maximum of 256 tokens). Finally, we take the first hidden state of the last layer of the transformer as input to a randomly initialized final linear classifier, and fine-tune all the parameters. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 5e-6. The fine-tuning is conducted during 10 epochs.

Results Table 3 presents the average best validation accuracy after a ten-fold cross validation. The best result is 94.26% after training on the Literary books corpus. We also conducted a cross-corpus training-validation experiment to see the classifiers' generalization performance. In Table 4, when we observe by row, testing on each validation corpus gets the best accuracy after training on the same corpus genre (the numbers on diagonal are in bold). When we observe by column, the best validation result for each training corpus is obtained by validating on the Literary books corpus (the numbers in the last line are underlined). Since the BLEU and CHRF3 value are the lowest for the Literary books corpus, this suggests that the differences between human and machine translations are the largest for this corpus, thus the models trained on different corpus genres could still get high performance on the literary corpus. It is worth noting that the standard deviation is the largest for OpenSubtitles corpus. A quick manual checking shows that OpenSubtitles corpus contains sentence alignment noise, which also explains why its BLEU and CHRF3 scores are quite low. In brief, the performance of human-vs-machine translation classifier is encouraging, which allows us to continue the following experiments.

3.2 Correlation with non-literal translation

Upon training this human-vs-machine translation classifier, we would like to verify the following hypothesis: is there a positive correlation between the prediction probability of human translation and the non-literal translations' proportion? It means that for each sentence, if the proportion of non-literally

[5]Eleven sentence-aligned EN-FR literary books are used: *Pride and Prejudice, Jane Eyre, Alice's Adventures in Wonderland, Moll Flanders, Robinson Crusoe, A Study in Scarlet, The Great Shadow, The Hound of the Baskervilles, Rodney Stone, Three Men in a Boat, The Fall of the House of Usher.* The corpus is originally available from `http://farkastranslations.com/bilingual_books.php`

[6]`https://github.com/pytorch/fairseq/tree/master/examples/translation`

[7]It's important to construct this training dataset where human and machine translations share the same source sentences. Otherwise, training a classifier using corpus from different origins (e.g. Europarl for human translations; Amazon product reviews for machine translations) could lead to an accuracy as high as 97%, however, the bias is significant due to the obvious differences on the average sentence length, the corpus domain, etc.

[8]`https://github.com/glample/fastBPE`

	Europarl	OpenSubtitles	TED Talks	Literary books
BLEU value	42.07	24.55	40.11	15.91
CHRF3 value	65.52	48.55	61.72	41.97
Average with standard deviation	84.55% ± 0.57%	78.86% ± 5.49%	73.92% ± 1.58%	**94.26%** ± 1.70%

Table 3: Human-vs-machine translation classification: average best validation accuracy after a ten-fold cross validation

Training 90% / Validation 10%	Europarl	OpenSubtitles	TED Talks	Literary books
Europarl	**84.55%** ± 0.57%	67.57% ± 3.30%	74.88% ± 1.97%	71.18% ± 3.26%
OpenSubtitles	72.55% ± 5.98%	**78.86%** ± 5.49%	72.84% ± 6.59%	71.76% ± 7.10%
TED Talks	66.44% ± 2.43%	64.54% ± 1.98%	**73.92%** ± 1.58%	62.29% ± 2.60%
Literary books	88.43% ± 2.35%	83.17% ± 4.77%	86.39% ± 2.47%	**94.26%** ± 1.70%

Table 4: Best average cross-corpus validation accuracy after a ten-fold cross-validation. The best results for each row are in bold, those for each column are underlined

translated tokens is high, we expect that the human-vs-machine translation classifier classifies the sentence as a human translation with a higher probability.

In order to save a final classifier model trained on all available data, after estimating the performance with the above cross-validation, we fix the number of epochs for which the average validation accuracy among ten folds was the highest. The chosen number of epochs are 5 for Europarl, 2 for OpenSubtitles, 9 for TED Talks and 8 for Literary books.

Next, we verify our hypothesis by applying the final classifier models on the corpus produced by Zhai et al. (2019), referred to as `TED-TT` henceforth. It's a subset of EN-FR TED Talks corpus manually aligned and annotated with translation techniques.[9] The inter-annotator agreement Kappa (Cohen, 1960) between two annotators on a control corpus is 0.67, which surpasses the substantial agreement threshold 0.61. This corpus contains 1 724 sentence pairs (37k English tokens and 38k French tokens), which are not included in our TED Talks training corpus. The translations of TED Talks' subtitles were generated by worldwide volunteers, instead of by professional translators as is the case for formal publication. For this reason, it is not the most favorable context for producing complex non-literal translations. Nonetheless, Zhai et al. (2019) still found a significant amount of examples in this corpus during the annotation. Therefore, it is worth it to later extend this work on literary translations.

In each sentence, the non-literal translations' proportion is the number of English tokens annotated with non-literal translation techniques divided by the total number of English tokens. In this experiment, the categories *literal, equivalence, lexical_shift* are combined together as literal translation; *transposition, generalization, particularization, modulation, modulation+transposition, figurative* are combined together as non-literal translation.[10]

Table 5 shows the Spearman's rank correlation coefficient (Hauke and Kossowski, 2011), which evaluates whether the rankings generated from the prediction probability of being human translation (obtained after applying softmax) are similar to the rankings generated from the non-literal translations' proportion. Using the model trained on Literary books corpus, we obtain the highest correlation coefficient 0.42, which is moderately positive and statistically significant (with p-value $< 1\%$).

3.3 Detecting the presence of non-literal translations

After obtaining a moderate correlation between the prediction probability of being human translation and the non-literal translations' proportion, we move our attention back to our first goal: detecting the

[9] Four computational linguists participated in the annotation, who are French and Chinese native speakers.

[10] The detailed definitions of each category are listed in the article of Zhai et al. (2019).

Human-vs-machine translation classifier trained on all available data	Spearman's ρ	P-value
Literary books	**0.42**	5.8e-73
Europarl	0.40	4.3e-66
OpenSubtitles	0.33	9.6e-44
Ted Talks	0.23	5.1e-23

Table 5: Investigate the correlation between the prediction probability of being human translation and the non-literal translations' proportion

presence of non-literal translations in a sentence.

To this end, we separate the TED-TT corpus in two classes: 1) 467 sentences containing only these categories: *literal, equivalence, lexical shift*; 2) 1 257 sentences containing these non-literal translation techniques: *transposition, generalization, particularization, modulation, modulation+transposition, figurative*. Since the dataset is imbalanced, the majority vote baseline is 72.91%.

We conduct a ten-fold cross validation on this dataset and we compare two approaches as presented in Table 6. Because we have observed a moderate correlation between human translation and non-literal translation (cf. Table 5), our hypothesis is that after a first fine-tuning of XLM to distinguish human and machine translation, loading this model to resume the fine-tuning task on detecting the presence of non-literal translations in a sentence could result in a better performance than directly fine-tuning XLM on the latter task.

The hyperparameters are the same as those used to train the human-vs-machine translation classifier. The results show that directly fine-tuning XLM on this dataset obtains a better accuracy (78.66%) than the majority vote baseline. Compared to this, 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 (80.16% and 79.86%, respectively).

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

Table 6: Detecting the presence of non-literal translations in a sentence. The best performance is in bold, and the second best is underlined

4 Detecting non-literal translations at phrase level

As shown in Table 1, non-literal translations occur more often at sub-sentential level, and are characterized by various translation techniques (Vinay and Darbelnet, 1958). Therefore, our upcoming research goal is to adapt this fine-tuning architecture to classify literal and non-literal translation at phrase level. This step is implemented by following the work of Arase and Tsujii (2019). Different from the efforts to pre-train larger models by giving enormous corpora for improvement (Liu et al., 2019; Yang et al., 2019), Arase and Tsujii (2019) proposed to inject semantic relations between a sentence pair into a pre-trained BERT model (Devlin et al., 2019), through simultaneous classification of sentential and phrasal paraphrases. Their phrasal paraphrase classification aims to give explicit supervision of semantic relations among phrases in sentence representation learning. Their work improved the sentence repre-

sentation learning, which is the basis for the central problem of assessing semantic equivalence in natural language understanding.

Preprocessing and architecture Inspired by the work of Arase and Tsujii (2019), we use the architecture presented in Figure 1, where the original entire sentence pair is as follows:

This, I think, is so deeply embedded in the water supply that it wouldn't occur to anyone to question it. $\rightarrow$ *Ceci est tellement intégré dans la pensée que personne ne penserait à le remettre en cause.*
(Lit. This is so ingrained in our thinking that no one would think to question it.)

After applying lowercasing, accent removing and BPE sub-word tokenization, the pair becomes:[11]

this , i think , is so deeply em@@ bed@@ ded in the water sup@@ ply that it would n't occu@@ r to anyone to question it . $\rightarrow$ *ceci est t@@ ellement integre dans la pensee que personne ne pen@@ serait a le re@@ mettre en cause .*

A special token `</s>` is put at the beginning and the end for both processed source sentence s and target sentence t. These two sequences concatenated (in form of tensors) serve as input to the XLM model. The dataset `TED-TT` provides the token indexes of aligned phrases for each sentence pair. The original alignment indexes $\{< (s_1, s_2, ..., s_n), (t_1, t_2, ..., t_m) >\}$ of each phrase pair are adjusted after all the preprocessing steps.

We fine-tune the same XLM's pre-trained model *mlm_tlm_xnli15_1024.pth*, by conducting the feature engineering as illustrated in Figure 1. The XLM model encodes the pre-processed sentence pair, and we obtain the final hidden states for each sub-word (i.e. output of the bidirectional Transformer, which captures rich contextualized features).

We then combine the corresponding hidden states to generate representations for source and target phrases (h_s and h_t), according to the adjusted phrase alignment indexes. We choose max-pooling as the combination function, which means selecting the maximum value over each dimension of the hidden units (Collobert and Weston, 2008; Conneau et al., 2017). The representations h_s and h_t are converted to a single vector to extract relations between them, by using these three matching methods: concatenating, calculating the absolute element-wise difference and taking the element-wise product (Conneau et al., 2017). This final vector is fed into a fully-connected classifier and we fine-tune all the parameters. These procedures are implemented in the forward function of our neural classifier.

Since most sentence pairs contain at least one aligned phrase pair, each sentence pair corresponds to a sequence of binary labels: literal or non-literal translation. This is different from our fine-tuning XLM at sentence level, where each sentence pair corresponds to only one label: human or machine translation (or containing or not non-literal translations). To implement this change, we modify the data loading and loss calculation function of our neural classifier.

Experiment The `TED-TT` dataset was annotated with multiple labels of translation techniques, and here we conduct binary classification by combining the labels as in section 3.2. However, the number of literal translations (21 791) is much larger than that of non-literal translations (2 234). In order to solve this problem, we randomly retain as many literal translation as non-literal translation examples for each sentence pair. This approach leads to an approximately balanced distribution: 2215 phrasal literal translation examples and 2234 phrasal non-literal translation examples, contained in 1110 sentence pairs.

We conduct a ten-fold cross validation on this dataset. The Adam optimizer is used with a learning rate of 5.5e-5. The fine-tuning is conducted during 10 epochs. The best average validation accuracy is 83.35%$\pm$1.49%, which is significantly better than the majority vote baseline 50%.

Evaluation A separately annotated test corpus (170 sentence pairs) is used for evaluating the final model trained with all available data (saved after 7 epochs). After the preprocessing steps as described above to obtain a balanced dataset, 100 sentence pairs containing 175 literal pairs and 177 non-literal pairs are kept. The test accuracy is 85.20% (300/352 pairs correctly predicted).

5 Comparison with state of the art and discussion

For our detection task at phrase level, we present the confusion matrix in Table 7. Among the 41 non-literal translations predicted as literal, the error ratio for each non-literal translation technique is as fol-

[11]The symbol '@@' is automatically added by the tool fastBPE after BPE tokenization.

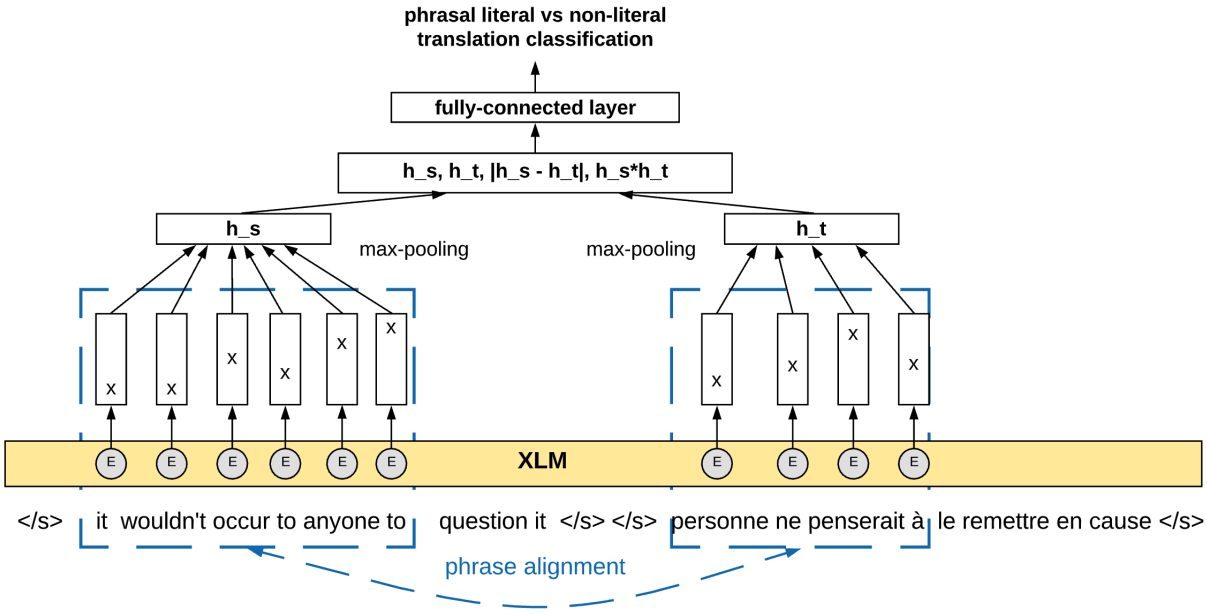

Figure 1: Fine-tune XLM at phrase level to classify literal vs non-literal translation. We omit the sub-word splitting of the input sentences for simplifying the presentation. For each pair of sentences, we conduct feature engineering for all manually aligned phrase pairs. For example, the pair *it wouldn't occur to anyone to* → *personne ne penserait à* (Lit. no one would think to) is a non-literal translation example; whereas the pair *question it* → *le remettre en cause* is a literal translation example

lows: 14/56 transposition, 11/31 particularization, 8/48 modulation, 5/31 generalization, 2/3 figurative and 1/8 modulation+transposition. And here is the distribution to explain the 11 literal translations predicted as non-literal: 4/8 equivalence, 6/160 literal, 1/7 lexical_shift. As an illustration, we present the definition of each translation technique and the corresponding prediction error examples in Table 9.

Gold \ Prediction	literal	non-literal
literal	**164**	11
non-literal	41	**136**

Table 7: XLM-based classifier (test accuracy 85.20%)

Gold \ Prediction	literal	non-literal
literal	**147**	28
non-literal	4	**173**

Table 8: Linguistic features-based RandomForest classifier (test accuracy 90.90%)

In order to compare the performance of this XLM-based classifier with a non-neural classifier which leverages linguistic features, we reuse the 198 features designed by Zhai et al. (2019). After a ten-fold cross validation on the same training data, a tuned RandomForest classifier[12] gives the best average accuracy of 82.91%±1.28%. We then apply the saved RandomForest model trained on all available data on the same test set, and get an accuracy of 90.90%. The confusion matrix is shown in Table 8.

The two confusion matrices show that the XLM-based classifier is better at detecting literal translations, whereas the linguistic features-based classifier is better at detecting non-literal translations. An oracle study shows that if we could unite the strength of both, only 7 wrong cases remain (1 non-literal translation predicted as literal, and 6 literal translations predicted as non-literal), which will result in an accuracy of 98.01%. Therefore, a hybrid classifier could be investigated in future work to improve the performance.

[12]The main hyperparameters are: n_estimators=1000, max_depth=50, min_samples_leaf=1, min_samples_split=3.

Translation technique	Error examples of the XLM-based classifier
Non-literal predicted as literal	
Transposition	Change grammatical classes without changing the meaning. *when I was a teenager → à l'adolescence* (Lit. in adolescence)
Particularization	The source word or expression could be translated into several target words or expressions with a more specific meaning, and the translator chooses one of them according to the context. *was met with this → ai été confrontée au* (Lit. was confronted with the)
Generalization	Several source words or expressions could be translated into a more general target word or expression, and the translator uses the latter to translate. *is going to ascend → arrivera* (Lit. will arrive)
Modulation	Metonymical and grammatical modulation; change the point of view; the meaning could be changed. *lifted my face up → ai levé la tête* (Lit. raised my head)
Figurative	Introduce an idiom to translate a non-fixed expression, or a metaphorical expression to translate non-metaphor. *putting everything I have → mets tout mon cœur* (Lit. put all my heart)
Modulation+Transposition	This category combines the transformations of Modulation and Transposition. *is no longer a glimpse of God → ne reflète plus Dieu* (Lit. don't reflect God anymore)
Literal predicted as non-literal	
Literal	Word-for-word translation, also concerns lexical units in multiword form. *get up → me lever* (Lit. get up)
Equivalence	Fixed translation of proverbs or fixed expressions; a word-for-word translation makes sense but the translator expresses differently, without changing the meaning and the grammatical classes. *fear-based → apeurée* (Lit. frightened)
Lexical_shift	The translation is not literal, but there is no change in meaning. They are minor lexical level changes, which do not involve any translation technique. *we → on* (Informal usage of "nous" (we))

Table 9: Definition of each translation technique and the corresponding error examples

6 Conclusion and perspectives

In this paper, we concentrate on the detection of non-literal translations by fine-tuning pre-trained cross-lingual language models (XLM).[13] A subset of TED Talks corpus manually annotated with translation techniques is utilized to ensure that we deal with appropriate human non-literal translations.

We first train a human-vs-machine translation classifier with different corpus genres and we show that there exists a moderate positive correlation between the prediction probability of being human translation and the non-literal translations' proportion for a sentence. The presence of the translation techniques *modulation* and *modulation_transposition* tends to make the classification as a human translation more difficult. For detecting whether a sentence contains non-literal translations, resuming the fine-tuning after loading the final trained human-vs-machine translation classifier brings a gain of performance (best accuracy 80.16%) than directly fine-tuning XLM. After adapting the architecture, our XLM fine-tuning at phrase level to distinguish literal and non-literal translations obtains an accuracy of 85.2%. We con-

[13]The dataset and code are available at `https://github.com/YumingZHAI/nlt_xlm`.

ducted detailed error analysis and compared the XLM-based classifier with a linguistic features-based RandomForest classifier. 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.

The automatic bilingual word alignment being a non-trivial task in itself (Song et al., 2020; Berrichi and Mazroui, 2020), our detection of non-literal translations at phrase level is currently based on already aligned pairs. In future work, it is important to leverage the advances of automatic alignment to reduce the reliance on manual work. To observe the generalization performance, one could extend these experiments to a more dissimilar language pair, for example English-Chinese. The fine-tuned models could be used to help constructing material to teach translation and to analyze the usage of different translation techniques between languages. We aim to automatically construct a corpus containing abundant non-literal translation phenomena based on this study, which we hope to be useful for the research on evaluating automatic word alignment, and for giving inspiration for MT's development to produce more appropriate non-literal translations.

Acknowledgements

We highly appreciate the reviewers' insightful comments and many detailed suggestions. We thank the French platform Saclay-IA for permitting the experiments conducted on its GPU cluster Lab-IA. The first author works as a postdoctoral researcher at Beijing Foreign Studies University, and we are grateful for the financial support given by the BFSU Artificial Intelligence and Human Languages Lab.

References

Lars Ahrenberg. 2017. Comparing machine translation and human translation: A case study. In *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*, pages 21–28, Varna, Bulgaria, September. Association for Computational Linguistics, Shoumen, Bulgaria.

Yuki Arase and Jun'ichi Tsujii. 2019. Transfer fine-tuning: A BERT case study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5393–5404, Hong Kong, China, November. Association for Computational Linguistics.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan, June. Association for Computational Linguistics.

S. Berrichi and A. Mazroui. 2020. Enhancing machine translation by integrating linguistic knowledge in the word alignment module. In *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pages 1–6.

Michael Carl and Moritz Jonas Schaeffer. 2017. Why translation is difficult: A corpus-based study of non-literality in post-editing and from-scratch translation. *HERMES-Journal of Language and Communication in Business*, (56):43–57.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79, Vancouver, August. Association for Computational Linguistics.

Qi Chen, Oi Yee Kwong, and Jingbo Zhu. 2018. Detecting free translation in parallel corpora from attention scores. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong, 1–3 December. Association for Computational Linguistics.

Hélène Chuquet and Michel Paillard. 1989. *Approche linguistique des problèmes de traduction anglais-français*. Ophrys.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, New York, NY, USA.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, October-November. Association for Computational Linguistics.

Dun Deng and Nianwen Xue. 2017. Translation divergences in chinese–english machine translation: An empirical investigation. *Computational Linguistics*, 43(3):521–565.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Bonnie J Dorr, Lisa Pearl, Rebecca Hwa, and Nizar Habash. 2002. Duster: A method for unraveling cross-language divergences for statistical word-level alignment. In *Conference of the Association for Machine Translation in the Americas*, pages 31–43, Berlin, Heidelberg. Springer.

Amel Fraisse, Zheng Zhang, Alex Zhai, Ronald Jenn, Shelley Fisher Fishkin, Pierre Zweigenbaum, Laurence Favier, and Widad Mustafa El Hadi. 2019. A sustainable and open access knowledge organization model to preserve cultural heritage and language diversity. *Information*, 10(10):303, Sep.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mengnan Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *ArXiv*, abs/1803.05567.

Jan Hauke and Tomasz Kossowski. 2011. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87–93.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical MT. In *Proc. ACL:Systems Demos*, pages 177–180, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Ekaterina Lapshinova-Koltunski and Christian Hardmeier. 2017. Discovery of discourse-related language contrasts through alignment discrepancies in English-German translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 73–81, Copenhagen, Denmark, September. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lucía Molina and Amparo Hurtado Albir. 2002. Translation Techniques Revisited: A Dynamic and Functionalist Approach. *Meta*, 47(4):498–512.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Brussels, Belgium, October. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ellie Pavlick, Johan Bos, Malvina Nissim, Charley Beller, Benjamin Van Durme, and Chris Callison-Burch. 2015. Adding semantics to data-driven paraphrasing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1512–1522, Beijing, China, July. Association for Computational Linguistics.

Minh Quang Pham, Josep Crego, Jean Senellart, and François Yvon. 2018. Fixing translation divergences in parallel corpora for neural MT. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2967–2973, Brussels, Belgium. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Spencer Rarrick, Chris Quirk, and Will Lewis. 2011. MT detection in web-scraped parallel corpora. In *Proceedings of MT Summit XIII*, pages 422–430. Asia-Pacific Association for Machine Translation, September.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

K. Song, X. Zhou, H. Yu, Z. Huang, Y. Zhang, W. Luo, X. Duan, and M. Zhang. 2020. Towards better word alignment in transformer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1801–1812.

Margita Šoštarić, Christian Hardmeier, and Sara Stymne. 2018. Discourse-related language contrasts in English-Croatian human and machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 36–48, Brussels, Belgium, October. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey, May. European Languages Resources Association (ELRA).

Antonio Toral and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment*, pages 263–287. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Jean-Paul Vinay and Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais: méthode de traduction*. Bibliothèque de stylistique comparée. Didier.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018. Identifying Semantic Divergences in Parallel Text without Annotations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1503–1515. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

Yu Yuan and Serge Sharoff. 2020. Sentence level human translation quality estimation with attention-based neural networks. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1858–1865, Marseille, France, May. European Language Resources Association.

Yuming Zhai, Pooyan Safari, Gabriel Illouz, Alexandre Allauzen, and Anne Vilnat. 2019. Towards Recognizing Phrase Translation Processes: Experiments on English-French. *CoRR*, abs/1904.12213.

Yuming Zhai. 2019. *Reconnaissance des procédés de traduction sous-phrastiques : des ressources aux validations (Recognition of sub-sentential translation techniques : from resources to validation)*. Thèse de doctorat, Université Paris-Saclay, December.