Shopee Code League 2020

Competition #4 - Title Translation FAQ

Dataset

Questions	Answers
Do I need to translate both the product title and category?	You only need to translate the product title.
Do I need to include the brand name during translation?	The brand name is important, please include it during translation.
Does each row in the training data Chinese csv match with the row in the English csv?	No, the training data Chinese csv does not necessarily match with the English csv. This is an unsupervised machine translation problem. You are required to solve it without parallel training data.
Is it guaranteed that each data item has a corresponding English pair?	No, it is not guaranteed.
Is the English csv and Chinese csv not in order or are they from two different data sets?	Both csvs are randomly pulled data from the Taiwan and Philippine markets.
Am I supposed to form the English - Chinese pairings from the csv files given or use readily available ones on the web?	Depending on which methods you are using, some may directly use the training data provided whereas others may choose to start from a pre-trained model.
Is the English csv for reference after translation?	No, it is not. Both the English and Chinese csv are training data.
The categories for the training and test data are different. Will it affect the training results?	The category names are different since they are from different markets (Taiwan and Philippines). The product category is extra information in case some teams find it useful. The categories should contain similar products.
Does the order of the words matter during the translation?	Yes, it does.
Will there be noise in the categorisation for the English and Chinese training dataset?	The category names may be different since the datasets are from different markets (Taiwan and Philippines). It is also not guaranteed that you will be able to find an exact translation in English for the Traditional Chinese title.

	Please note that both training datasets are not parallel.

Special Characters

Questions	Answers
What do I do about the special characters such as emojis, stars, etc.	You can ignore the special characters. They are not counted during evaluation and will not affect your scores.
I noticed that there are some English words mixed in the Traditional Chinese text, would there be issues in tokenization?	No, there would not be issues.

Language Concerns

Questions	Answers
What are some language-aware tokenizers I can use?	Some common tokenizer for Chinese is Jieba and for English is white space or moses tokenizer.
Would it be a disadvantage if I cannot understand traditional Chinese?	The impact should be minimal. The problem statement requires your understanding and methods of tackling the task, rather than the ability to understand the language.
Since many of the Chinese-English translation corpora are in Simplified Chinese, what would be the fastest way to convert to Traditional Chinese?	You can use open-source tools such as opencc.
Should I be concerned about the Chinese set being in Traditional or Simplified Chinese?	As long as you have a consistent input, there should not be a concern.
Since Chinese title translation to English title translation may vary between submissions, how will the judging be like?	We have a human reference file and it is evaluated with sacrebleu.

How can I check the translation quality?	We will release a validation set soon, so you can refer to the score on the validation set to judge your model performance
Some Traditional Chinese titles contain English words. Is that done on purpose?	No, it is not. English brands are more commonly used when sellers label their products. These are raw titles from actual sellers.

Noisy Data

Questions	Answers
Can I manually clean the data?	Yes, you can manually clean the data. However, you cannot hand label the test set.
Can I remove the punctuations and emojis when cleaning the data?	Yes, you can remove the punctuations and emojis.

Models

Questions	Answers
Can I use a pre-trained model?	Yes, you can
Can I use open source code from github?	Yes, you can.
Am I allowed to use Google Translate's API?	No, you cannot directly use Google Translation and submit the result. This is strictly not allowed.
What are some recommended algorithms for translation models?	You can search for github open-source code from Facebook's unsupervised machine translation.

External Data

Questions	Answers
Am I allowed to scrap data from Shopee's website for more titles?	Yes, you can. However, please avoid scraping too hard.
Can I use other public datasets?	Yes, you can. The use of public data or any pre-trained model is encouraged.

Submission

Questions	Answers
Will there be a sample of the evaluation csv file?	We will release a sample submission shortly before the evaluation platform is launched.
Are punctuations, brackets, slash, etc. required in the final translated output?	You can include them, but it will not be considered during the evaluation.
How many submissions per day?	There is no per day limit but you can only make a total of 15 submissions for this competition.
During the submission of the model, am I submitting the entire jupyter notebook?	Yes, please submit the complete scripts with any dependencies specified - which should output the .csv file used for submission.

Scoring

Questions	Answers
How does BLEU scoring work?	You can check the github page here, https://github.com/mjpost/sacrebleu
How is the translated output graded?	It will be graded based on professional human translators' translation.

Others

Questions	Answers
The Traditional Chinese dataset is missing a row of entry, in row 490233. Is that row empty on purpose?	No, it is not empty on purpose. The data was pulled from the Shopee platform, so the product may have an empty title. You can just ignore it.