Technical Disclosure Commons

Defensive Publications Series

September 2021

Number Translation and Unit Conversion Using Machine Learning

Haijie Hong

Lu Sun

Number Translation and Unit Conversion Using Machine Learning

ABSTRACT

Machine translation is widely utilized to translate text between different language pairs. Applications of automatic translation include content localization. Different regions of the world utilize different measurement units (e.g., acre vs. hectare). Correctly converting and translating measurement units is thus an important part of content localization. Current machine translation models have low accuracy when translating numbers and are unable to handle unit conversions. This disclosure describes techniques to train a machine learning model such that it can generate accurate translations of numbers, including unit conversions. A base model is trained using input text that is tokenized, including splitting numbers into individual digits. Parameters of the trained base model are used to initialize a custom model that is fine-tuned using training data that has been augmented to include annotations, e.g., different values and units for each measurement in the source text. The trained custom model described can deliver correct number translations and unit conversions and can be used for content localization.

KEYWORDS

- Machine translation
- Number translation
- Unit conversion
- Content localization
- Neural translation
- Translation model
- Data augmentation
- Data annotation

BACKGROUND

Machine translation is widely utilized to translate text between different language pairs. Applications of automatic translation include content localization. Different regions of the world utilize different measurement units (e.g., acre vs. hectare). Correctly converting and translating

measurement units is thus an important part of content localization. Current machine translation models have low accuracy when translating numbers and are unable to handle unit conversions. Some examples of incorrect unit conversion in machine translation are presented below (the text in red indicates the incorrectly translated/converted portion of the text). The consequence of such errors is significant since the values in the translated text are incorrect.

English Source Text	**Incorrect German Translation**
Popular street known for its 2-acre nature reserve & verdant green park for families & wildlife.	Beliebte Straße mit 2 ha großem Naturschutzgebiet und grünem Park für Familien und Wildtiere.
260 ft. in length, this road is the city's narrowest street & dates back to the 17th century.	Die engste Straße der Stadt, die bis ins 17. Jahrhundert zurückreicht, ist 25 m lang .

Translation and content localization is commonly performed using machine learning models. The problems with number translation/conversion can include:

- Digit instability where digits are deleted, inserted, replaced, or out-of-order in the translated text.
- Numbers off by 10x, especially during number format conversion. This is often a problem when translating between languages with different number representations.
- Unit conversion errors occur when a translation model generates incorrect numbers in the translated text when converting to local units.
- Semantic confusion occurs when there is a series of numbers and the translation model fails to understand the semantics which results in incorrect translation.

DESCRIPTION

This disclosure describes techniques to improve the quality of unit conversion in machine translation models. The unit conversion problem is solved through an annotation approach. In the training data used to train the model, measurements in source text are annotated using a semantic extractor. Further, translations in target languages are provided with different measurement units. At the inference stage, the machine translation model picks an appropriate translation from the annotations. The custom machine translation model described in this disclosure can provide substantial improvement in the quality of machine translation of numbers as well as unit conversions.

Fig. 1 illustrates an example process for custom machine translation as per the techniques described in this disclosure.

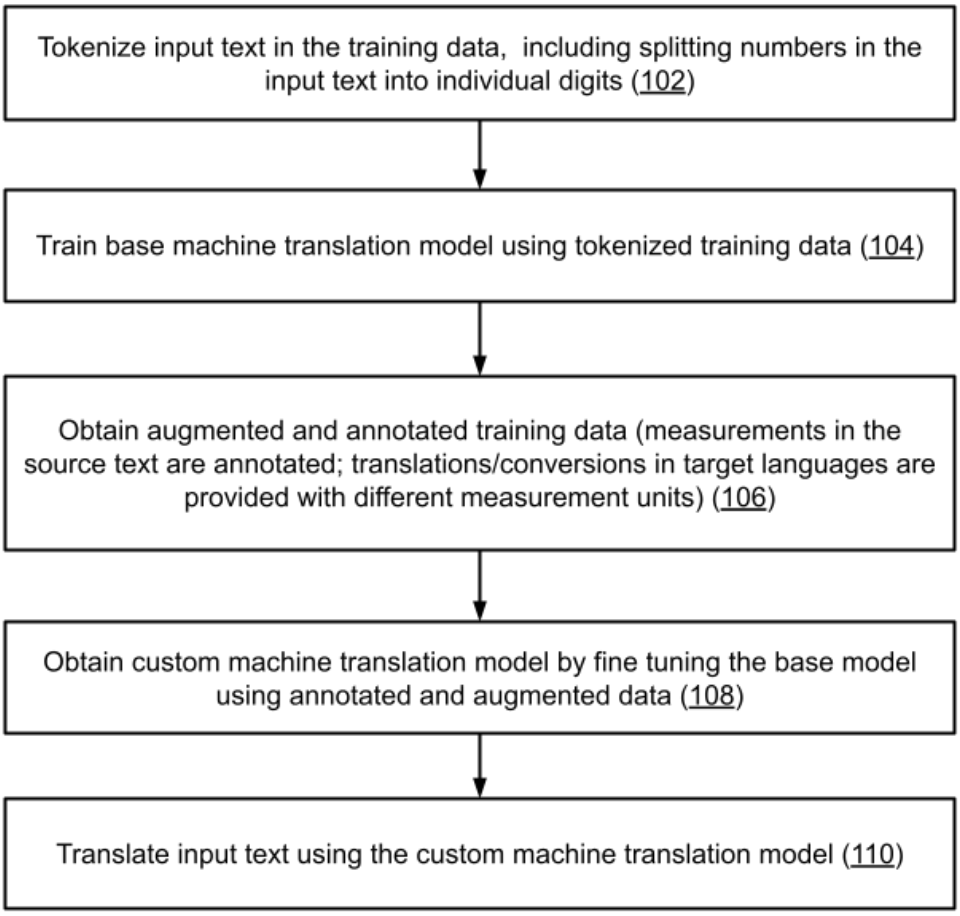

Fig. 1: Example process for custom machine translation with unit conversion

Fig. 1 illustrates an example process to train a machine translation model that provides high accuracy of number translation and unit conversion. Training data includes input text in a source language. Input text from the training data is tokenized for machine translation (102). During tokenization, numbers in the input text are tokenized into individual digits, e.g., using a split digit tokenizer. A base machine translation model is trained using the tokenized input data (104).

Additional training data is obtained and is augmented and annotated (106). Specifically, the text is augmented and annotated to include units that different target languages may use, and measurements in the source text are converted to accurate values in the various units. For example, the raw source text "Access to a 3-acre botanical display" is annotated to "Access to a 3-acre [12.140,57 m², 3 ac, 1,21 ha] botanical display." As seen in the example, the annotated text (in blue) includes 3 different representations of the number "3" and the associated unit "acre."

The base translation model is used to initialize the parameters of a new custom translation model that is incrementally trained with the additional training data (108). The annotated and augmented content in the training data enables fine tuning of the model via the incremental training process. The trained custom translation model can then be deployed to translate input text (110). For example, the model correctly outputs the German text "Zufahrt zu einem 1,2 ha großen Park" as a translation of the example sentence above. In this case, the translation includes unit conversion from 3 acres to 1.2 hectares, as the appropriate unit of measurement for German is hectare, while that in the source English text is acre.

In this manner, the described techniques provide a machine translation model that has

high accuracy when translating numbers and converting between measurement units. The model can be utilized for content localization. In some cases, if the model is to be used in a specific domain (e.g., travel), the model can be trained using both out-of-domain data (generic sentence pairs) as well as in-domain data (sentence pairs from the travel domain). In this case, model fine tuning may be performed specifically with in-domain data.

CONCLUSION

This disclosure describes techniques to train a machine learning model such that it can generate accurate translations of numbers, including unit conversions. A base model is trained using input text that is tokenized, including splitting numbers into individual digits. Parameters of the trained base model are used to initialize a custom model that is fine-tuned using training data that has been augmented to include annotations, e.g., different values and units for each measurement in the source text. The trained custom model described can deliver correct number translations and unit conversions and can be used for content localization.

REFERENCES

1. Dinu, Georgiana, Prashant Mathur, Marcello Federico, Stanislas Lauly, and Yaser Al-Onaizan. "Joint translation and unit conversion for end-to-end localization." *arXiv preprint arXiv:2004.05219* (2020).