Smart Pre- and Post-Processing for STAR MT Translate

Judith Klein
STAR Group
Wiesholz 35, 8262 Ramsen
Switzerland
Judith.Klein@star-group.net

Abstract

After many successful experiments it has become evident that smart pre- and post-processing can significantly improve the output of neural machine translation. Therefore, various generic and language-specific processes are applied to the training corpus, the user input and the MT output for STAR MT Translate.

1 STAR MT Translate

STAR MT Translate is STAR's web-based MT system which can also be integrated in Microsoft Office products using the STAR MT Office Connector. In this MT application all data is kept safe within the customer's environment.

The aim is to provide "useful" translations for customers in their company-specific domain, e.g. transportation. Typically, the users are not translators but various professionals, e.g. mechanics, who need to understand the information that is only available in a foreign language.

In professional translation projects mostly structured text is translated, possibly supported by MT. The user input to STAR MT Translate however is much more flexible and "unpredictable". It contains customer-specific terminology but it consists of a wide range of linguistic phenomena, including ungrammatical sequences.

To meet both kinds of text, firstly, the core of the engines are built from customer-specific in-domain translations; secondly, to deal with the variety of language usage, the engines are complemented by out-of-domain data, and they use neural technology.

2 Smart Pre- and Post-processing

Good training data is the essential requisite to obtain good MT. It must cover the language phenomena that are likely to occur in the user input sent to the MT system for translation.

Even if the corpus includes the required characteristics it usually also contains "noise" that considerably reduces the quality of the MT output.

Therefore, STAR has developed a systematic strategy to identify and delete this "noise".

Firstly, language-independent processing rules delete incomplete formatting or punctuation, irrelevant characters, fragmentary sentences etc. Nominalizations of various, inconsistent number formats (dates, decimal separators, etc.) as well as URL and email addresses are defined in another step. A specific set of rules ensures that the corpora contain a balanced amount of similar sentences (regarding the length of sentences, the number of tokens, etc.) and determine the prioritization of segments depending on their completeness.

Finally, language-specific processes are applied that identify irrelevant text, e.g. typos, foreign language, informal expressions, and – depending on the language – include a special handling of morphological variants.

The same generic processing steps and the adequate language-specific rules are applied to the input text, in order to send to the MT engine sentences that are made up in the same way as the ones it has learned.

And finally, post-processing uses these rules, too, in order to obtain high-quality MT output.

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 355
Alacant, Spain, May 2018.