Combining different tools to build a semi-supervised data collection model to increase MT quality and performance

Short Abstract

Quality of Machine Translation (MT) output is clearly driven by the quality of the input. Those LSPs, like Capita, in search of the best way to implement MT within their industry/domain specific workflows, are continuously looking for mechanisms to assist them in the collection of quality corpus that can be used for engine building and training.

Our NLP team have been working during the past 3 years on building and training a number of "non-customer" domain specific engines for different language pairs. During this process they have built a semi supervised corpus collection tool that has helped to significantly reduce collection time and increase the quality of the data.

We would like to share our model and explain how we have used existing tools in combination with our in-house developments to build a data collection model that uses keywords generated from domain specific material to assess compatibility with the domain. We will then use these keywords to identify compatible monolingual data, which will then be paired to target languages. Once bilingual data is available, our tools will produce segmented bilingual corpora using alignment tools.

Finally, we reach a set of data that can be evaluated and cleaned both using technology and human evaluation processes.

Long Abstract

This is a study to combine a number of existing technologies with newly developed tools to create a semi-supervised tool to assist with corpus collection for MT. This study aims to combine technologies for domain classification, domain source identification, and comparable file alignment into a unified tool. The unified tool will be used to make the corpora collection process more focused and efficient enabling a wider variety of sources to be used.

- **Domain classification:** This is a process using domain specific material to generate lists of domain specific keywords. The domain specific material will simply be a set of plain text files which are known to have content related to the domain. The basis for this section of the tool will be topic classification technologies, which will be used to provide a list of keywords that have been ranked by their compatibility to the domain.

- **Domain source identification:** This section of the tool will use the keywords, and combinations of the keywords, to drive web searches to generate URLs for websites that produce hits matching the keyword combinations. This list of URLs will be cleaned up to remove duplicate URLs. Sample content will also be generated from these URLs and this content will be processed with the topic classification tool to verify the extent to which the URL matches the domain, using the topic classification tools ranking system. The lowest ranking URLs will be discarded. This process may be used to provide additional keywords for improved and more focused Domain identification.

- **Candidate URL pairing:** The previous stage in the process will generate a list of monolingual URLs. The next stage is to determine if a matching URL exists with content written in the

target language. This will require tools that can determine if a matching URL exists based on the target language.Additionally, the tool will have to identify the language the matching URL is written in.

- **Comparable file alignment:** The content of the paired URLs is aligned using comparable file alignment techniques to produce segment based corpora. This process will divide the content, both source and target, into individual segments using standard text segmentation structures. The segments will then be aligned based on a simple language model and a number of string comparison techniques. The end result will be a comparability rating to each segment combination. The segment combinations that exceed the threshold will be exported for further processing, the remainder will be discarded.

- **Corpora evaluation and clean up:** The produced corpora is evaluated both mechanically and through the use of human linguists. Poor quality segments will be discarded. The file format sent to the linguists will be in a simple Excel spreadsheet andthe linguists will be asked to rate the segments on a basic scale. As there may be very large numbers of segments to be rated, the segments will be divided into batches based on the source URLs and the linguists will be asked to rate the segments at a batch level.

Mark Unitt

Mark initiated his career in the translation industry back in 1991 as a localization engineer. From early days, he was involved on the development of computer assisted translation environments, having developed methodology for IBM TM2, Trados and DejaVu systems.

In 2004, as part of an initiative by Applied Language Solutions, Mark started working with a team of NLP professionals in the development of Moses based Machine Translation environments and the application of Machine Translation as part of the Localization workflows.

Currently, Mark is Head of Language Product Development at Capita Translation and Interpreting, where he manages a team of language experts and software developers working on a number of projects to support the integration of technology in the translation workflows. These projects include areas such as predictive analysis of Machine Translation quality, automation of workflows, Big data analysis and data mining and collection for corpus creation.