Optimized version of Translation memory from Nynorsk Pressekontor (Nynorsk News Press Agency)

Vitec MV (Kirsti Skjervheim, Kristina Watt Aspmo, Stefanie Wolf)

Kontakt til sprøgsmål: Kristina.Aspmo@vitecsoftware.com eller stefanie.wolf@vitecsoftware.com

The automatically created translation memory of the Nynorsk Pressekontor has been optimized for use of statistical methods of natural language processing using the processes described below. Due to the scope of the optimization process, not all of the described steps have been applied on all of the corpus, this is pointed out as it might be of relevance in some applications.

Removal of translation units from duplicate or near duplicate articles

The goal of this procedure was to have a representative corpus that can be used for training statistical models of natural language. The parallelized corpus the translation memory is based on contains duplicate articles, near duplicate articles with varying nynorsk translations and articles that are expanded upon where both the long and the short versions exist.

The translation units, that are based on these articles have been reduced to the number of occurrences they appear in different contexts i.e. articles that are not different versions of the same article. In case of near duplicate articles, the best Norwegian Nynorsk translation is chosen. In case of the articles that are expanded upon, the translation units from the longest article are kept.

Note that this does not remove all duplicate translation units. If a sentence, typically a short and general one appears in several different articles, the number of occurrences of this sentence in the translation unit will correspond to the number of different contexts the sentence appears in.

Correction of sentence splits

Some of the translation units consisted of several sentences. In some cases, this is due to a missing space between sentence final punctuation and the first word in the following sentence.

We have fixed these errors and split translation units into single sentences as far as possible. The newly created units can be identified by the creationid=”Split”

Removal of non-aligned translation-units

Translation units where Norwegian Bokmål and Norwegian Nynorsk do not correspond to each other are removed from the translation memory. They are mainly removed as they have been encountered during manual processing.

Some of the nonaligned units have been identified in an automatic processing step. The translation units that consisted of more than one sentence were split. The cases where the number of bokmål and nynorsk sentences was unequal and one of the sentences did not have sentence final punctuation were collected automatically. The quality of this step was checked manually, and the method gave a good result containing very few errors.

The file **nonaligned-manually.xml** contains the proofread nonaligned units, while the file **nonaligned-automatically.xml** contains the automatically retrieved ones.

Other cases, where the number of bokmål and nynorsk sentences in a translation unit was unequal were marked with nonaligned. In some cases, the unequal number is due to the translation splitting one sentence up into two, in other cases it is due to an error in alignment. Even though this difference is very important, it was out of scope of this project to handle it. The segments are tagged with “nonaligned” in the translation memory.

As this work has not been done systematically there will still be non-aligned translation units in the translation memory.

Reconstruction of sentences

Some translation units in the original translation memory consist of one part of a whole sentence. This is where sentences from the original article have been split into two or more subparts, mostly because of the original article formatting. In other cases, a translation unit consists of several sentences because of a missing space after the final punctuation.

The translation units from the split sentences have been merged into a single translation unit. Translation units with several sentences have been split into one translation unit per sentence. The translation units created in this process are tagged as “Merged” in the creationid-attribute.

Removal of “non-language” data

Translation units such as addresses and captions with only names are removed as they occurred in the optimization process.

Markup of segments – Titles, captions and lists

Language use of article titles and captions is different from language use in running text.

We have tagged all of the segments that are the articles main titles in the translation memory. Captions, subtitles and list items have been tagged in so far as they occurred in other parts of the optimization process. See **Non-Sentences.md** for further details.

Markup of linguistic errors

Linguistic errors have been tagged as far as the language expert came across them in the optimization process. See **LinguisticErrors.md** for error types. The error types are coarse grained. Corrections can improve the quality of language data for use in statistical methods.

See **LinguisticErrors.md** for further details.