# Use MT to simplify and speed up your alignment for TM creation

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

**Giorgio Bernardinello**
STAR Group
Wiesholz 35, 8262 Ramsen
Switzerland
giorgio.bernardinello@star-group.net

## Abstract

Large quantities of multilingual legal documents are waiting to be regularly aligned and used for future translations. Due to time restraints and the effort and cost required, manual alignment is not an option. Automatically aligned segments are suitable for concordance search but are unreliable for fuzzy search and pre-translation. MT-based alignment could be the key to improving the results.

## 1 Why align if there is MT?

Today, in the translation sector, many texts and documents in different sectors and areas and in many different language combinations are translated using MT or with MT support. Why go to the effort of using alignment to create a Translation Memory (TM) for a CAT tool from existing translations? This is because precisely the existing verified translations are required. MT can probably provide precisely these translations by chance, but not reliably.

## 2 What data and for what purpose?

The Systematic Collection of Legislation (SR)[1] of the Swiss Federal Administration (Bund) is currently automatically aligned and provided as a TM[2] four times a year. It consists of over 5000 MS Word documents in German, French and Italian, each with over two million segments or almost 20 million words[3]. The final version of the legal text is synchronised or translated into all three languages in the original document.

Since the legal texts have been automatically aligned and not translated in the CAT tool, there is no TM with reliable segment alignment. The TM can therefore be used via concordance search[4] but is not suitable for fuzzy search or pretranslation.

## 3 STAR's translation technology

From 2020, the Bund will use STAR's translation technology as its general translation solution with the translation memory tool Transit as the core system. An alignment tool is integrated into Transit as standard. The MT interface makes it easy to use machine translations from different MT systems.

## 4 Interactive alignment in Transit

In alignment projects, the document pairs are imported in the respective languages and segmented independently of each other. The segment alignment then does not match, for example, if two sentences in one document correspond to just one sentence in the other document, if a sentence has no equivalent in the other text, or if the sentence order is different.

As usual with alignment tools, formal (e.g. sentence length, numbers, formatting) and lexical parameters (e.g. dictionary entries or unchanged words, such as company names) are used to calculate the segment alignment.

[1] https://www.admin.ch/gov/de/start/bundesrecht/systematische-sammlung.html

[2] Previously in MultiTrans.

[3] Version: March 2020

[4] In Transit, the user can open the aligned document pair in the Transit editor directly from the concordance result and copy the corresponding section of text. The same applies for fuzzy hits, but this would not be an efficient way of using the fuzzy search.

### 4.1 TM for the segment alignment

In addition to formal and lexical parameters, the TM can also be used to calculate the segment alignment. The Transit fuzzy algorithm[5] calculates the similarity between the target-language segment in the TM and the target-language segment in the aligned document. This value has the strongest weighting.

However, a TM that contains 100% hits is only available for an alignment project in exceptional cases, because, if there were a 100% hit for lots of segments, it would not be necessary to create another TM.

### 4.2 Alignment mode

The alignment tool uses the criteria to dynamically calculate the segment alignment. Change proposals are displayed graphically and in colour in the alignment editor. Manual checking and correction guarantees reliable segment alignment, meaning that the material can be used for the fuzzy search and pretranslation.

However, no matter how well the interactive alignment supports the user, interactive alignment alone is not an effective solution for such vast quantities of text as the SR.

## 5 Machine alignment in Transit

### 5.1 MT for the segment alignment

Sennrich and Volk (2010) have already reported on the successful use of MT for alignment, provided that the MT system produces good translations. This is not a question of the quality of the translation as such, but about its similarity to the existing translation, which is determined using "BleuAlign". BleuAlign is a part of BiTextor, which is used within the Paracrawl project to create a vast quantity of bilingual corpora.[6]

### 5.2 Machine alignment function

In the same way that the alignment tool uses the Transit fuzzy value from the TM translation for the segment alignment, it also uses the similarity value from the MT translation.

The selected MT system is specified in the alignment project. The following steps are automatically carried out when the "machine alignment" function is used:

- Machine translation of the source-language documents
- Evaluation of the MT similarity for the segment alignment
- Automatic modification of the segment alignment based on the results of the calculation

The changes are based on the segmentation of the source language and consist of (a) joining source- or target-language segments[7] in the case of 1:N or N:1 relationships, (b) inserting empty segments where there is missing target-language text[8], (c) deleting source- and target-language segments if multiple segments cannot be aligned, and (d) moving target-language segments.

### 5.3 Machine alignment of the SR

The machine alignment of the SR will be carried out automatically every three months via the "STAR CLM" workflow system[9]. In the subsequent runs, only changed or new texts are sent to the MT system (in this case, DeepL Pro[10]), since the other MT translations are already available. Whether and when the entire SR will be machine-translated again depends on whether innovations in the MT system could further improve the segment alignment.

## 6 Conclusion

The perfect interaction between the alignment tool and the MT interface in Transit means that machine translation can be directly integrated into the calculation of the segment alignment. This additional information can be used to improve the alignment, making both the

[5] Proven algorithm for the similarity calculation of source-language segments for pretranslation and fuzzy search.

[6] https://www.slideshare.net/TAUS/bitextor-harvest-your-own-parallel-corpora-from-the-web-miquel-esplgomis-universitat-dalacant, https://paracrawl.eu/

[7] To find out how "virtual join" works for the source language, see: https://www.star-group.net/en/downloads/transit-termstar.html

[8] Transit ignores a segment if the target is empty.

[9] The first cycle is planned for June 2020.

[10] https://www.bk.admin.ch/bk/de/home/dokumentation/medienmitteilungen.msg-id-77610.html

interactive and the automatic alignment easier and quicker.

We are still investigating whether machine alignment can be used to achieve reliable segment alignment in such a way that the material can be used not only for concordance search, but also for fuzzy search and pretranslation.

## References

Sánchez-Gijón, Pilar, Joss Moorkens and Andy Way. 2019. Post-editing neural machine translation versus translation memory segments. *Machine Translation* 33(1-2):31-59

Sennrich, Rico and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel text. *AMTA 2010. 9th Conference of the Association for Machine Translation in the Americas.* Denver, Colorado