STAR Transit & STAR MT: Morphologically Generated Additional Information for Improving MT Quality

Nadira HOFMANN

STAR Group, Wiesholz 35, 8262 Ramsen, Switzerland

nadira.hofmann@star-group.net

Abstract. As an experienced developer of language processing solutions, STAR Group applies its proven technologies to MT training: Transit's morphological support for 80+ languages is used to extract inflected terminology from dictionaries and reference material without any additional effort. This additional input for the engine training noticeably improves the MT quality – especially in morphologically rich languages.

Description

For STAR's MT system STAR MT, validated customer-specific dictionaries are used for MT training. However, dictionaries contain terms in canonical forms, while text corpora predominantly contains inflected forms. A large proportion of the terminological potential would remain untapped if engines were only trained with canonical forms.

This is where STAR's experience pays off: The TMS Transit offers morphological terminology support for over 80 languages -- not just simple stemming, but using linguistic expertise mapped out in rules that have been tried and tested throughout years.

As a result, inflected forms in the source and target language of a bilingual body of text are reliably identified and used to enrich the training material. This additional information is particularly valuable for the quality of the MT because it is validated twice: Customer-validated canonical forms that are retrieved from the dictionary along with their inflected forms that are actually used in the validated translation memory.

With this approach, STAR MT benefits from Transit's existing and proven morphological technology to create terminology that offers added value without investing any time. In practice, this brings about significant improvements in quality: For a customer-specific German-French engine, an additional 17% of terminology was extracted and the BLEU score increased by 1.1 points. The translations that are created from this were clearly selected as preferred translations by the translators who carried out the manual evaluation of the sentence BLEU lists.

References

Pinnis, M. (2015). Dynamic Terminology Integration Methods in Statistical Machine Translation, Proceedings of the 18th Annual Conference of the European Association for Machine Translation, 89-96.