Quoniam

Document Meta-Information as Weak Supervision for Machine Translation

Dr Laura Jehl (2018)

Abstract – Dissertation

Data-driven machine translation has advanced considerably since the first pioneering work in the 1990s with recent systems claiming human parity on sentence translation for highresource tasks. However, performance degrades for low-resource domains with no available sentence-parallel training data. Machine translation systems also rarely incorporate the document context beyond the sentence level, ignoring knowledge which is essential for some situations. In this thesis, we aim to address the two issues mentioned above by examining ways to incorporate document-level meta-information into data-driven machine translation. Examples of document meta-information include document authorship and categorization information, as well as cross-lingual correspondences between documents, such as hyperlinks or citations between documents. As this meta-information is much more coarse-grained than reference translations, it constitutes a source of weak supervision for machine translation. We present four cumulatively conducted case studies where we devise and evaluate methods to exploit these sources of weak supervision both in low-resource scenarios where no task-appropriate supervision from parallel data exists, and in a full supervision scenario where weak supervision from document meta-information is used to supplement supervision from sentence-level reference translations. All case studies show improved translation quality when incorporating document meta-information.

Read more: https://archiv.ub.uni-heidelberg.de/volltextserver/26780/1/Dissertation_Jehl_HeiDOK.pdf