Discourse-level Features for Statistical Machine Translation

The talk will show how the disambiguation of discourse connectives can improve their automatic translation. Connectives are a class of frequent functional lexical items that play an important role in text readability and coherence. Longer-range context is taken into account to learn the signaled rhetorical relations. The labels obtained from a discourse connective classifier are then integrated into statistical translation models (from EN to FR and DE). Linguistic annotation and evaluation issues are discussed together with results from automated scoring. Discourse connectives are furthermore especially prone to translationese and translated texts show either increased use or decreased use of discourse markers. Work in progress aims to capture this natural explicitation and implicitation of discourse connectives in current statistical machine translation models.

Along the same lines, the correct translation of verb tenses is important to maintain, in the target text, the same temporal ordering of events as in the source text. The utility of automatically labeling English Simple Past verbs is assessed with a binary discursive feature, narrative vs. non-narrative, for statistical machine translation into French. The narrativity feature helps deciding which of the French past tenses is a correct translation of the English Simple Past.