Evaluating Subtitle Segmentation for End-to-end Generation

Problem
Reference-based evaluation for perfect/imperfect texts

Per: He said, <eob> "It almost <eob> doesn’t matter what you know.”
Ref: He said, <eob> "It almost doesn’t matter <eob> what you know.”
Imp: He told me, <eob> "What you know <eob> is not important.”

Perfect: Precision=½ Recall=½
Imperfect: Precision? Recall?

Standard segmentation metrics cannot be computed for imperfect texts.
How to evaluate segmentation for imperfect texts?

Contributions
- Exp1: A comparison of sequence segmentation metrics for perfect texts
- Exp2: Sigma, a new segmentation score for imperfect texts
- Exp3: A boundary projection method to compute standard segmentation metrics for imperfect texts
- EvalSub: A tool for computing reference-based segmentation scores for automatic subtitles

Exp1: Metric sensitivity/robustness

-> Controlled segmentation degradations of reference

Precision-Recall for over-/under-segmentation

BLEU\textsubscript{br}
- balance among error types
- insensitive to shift size

Exp2: What does BLEU\textsubscript{br} really measure?

BLEU\textsubscript{br} largely depends on text quality (BLEU\textsubscript{nb}). The relative difference with BLEU\textsubscript{nb} cannot be used to measure segmentation quality in general.

Exp3: Boundary projection

Project boundaries from Hypothesis to Reference

Compute standard metrics

Sigma: a new segmentation metric

Sigma: ratio to a BLEU\textsubscript{br} upper bound, computed from the proportion of boundaries and n-gram precisions. Stable irrespective of BLEU\textsubscript{nb}

System ranking: NMT>Cas>e2e

Data and implementation

Data: MuST-Cinema test set EN→FR

Code available at: https://github.com/fyvo/EvalSubtitle

Conclusions

Sigma is a new BLEU-based metric suited for evaluation of subtitle segmentation in the case of end-to-end generation systems.
Future work: correlation with human judgements