# Evaluating Subtitle Segmentation for End-to-end Generation Systems

Alina Karakanta, François Buet, Mauro Cettolo, François Yvon

FONDAZIONE _ _

UNIVERSITY OF TRENTO - Italy

HUMANE AI NET

## 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 <eol> what you know."

**Imp**: He told me, <eob> "What you know <eol> is not important."

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

⇨ Standard segmentation metrics cannot be computed for **Imperf**

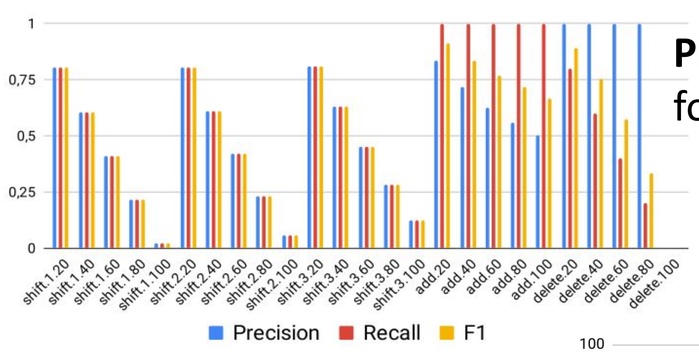
**How to evaluate segmentation for Imperfect texts?**

## 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

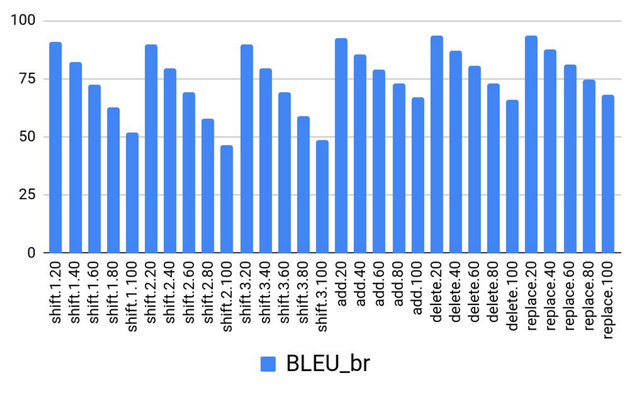
=> Controlled segmentation degradations of reference




**Precision-Recall**
for over-/under-segmentation

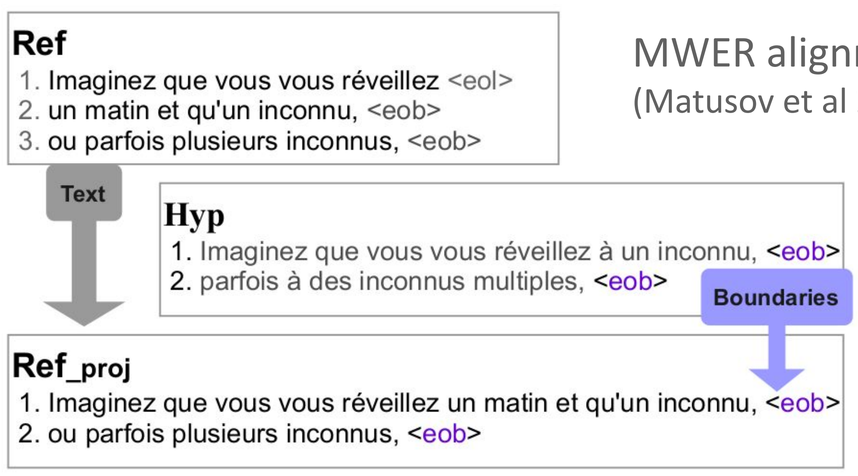
**BLEU$_{br}$**

- balance among error types
- insensitive to shift size




## Exp3: Boundary projection

Project boundaries from Hypothesis to Reference
Compute standard metrics

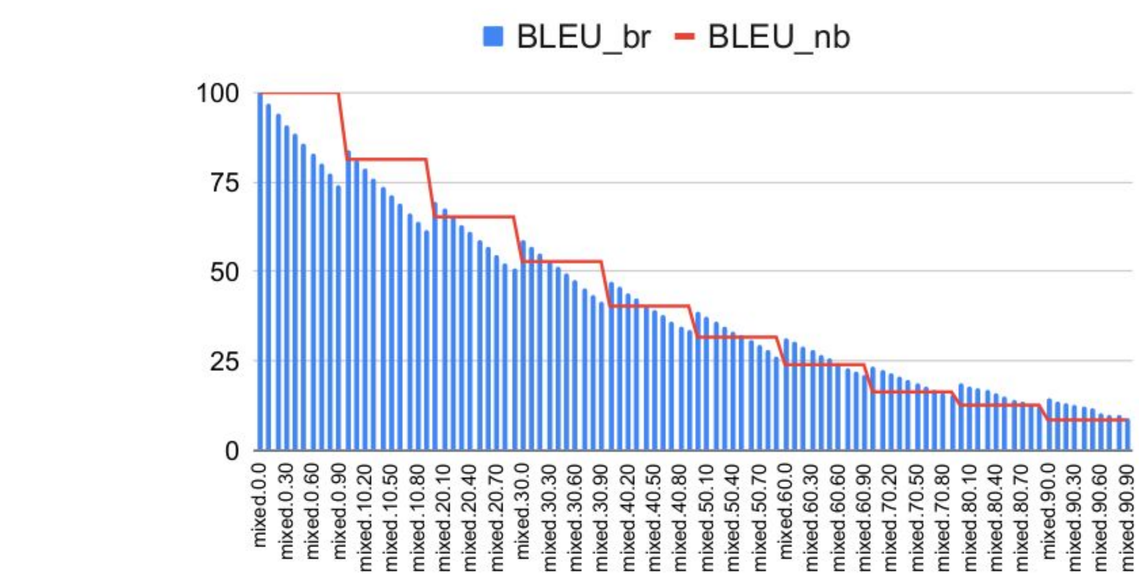



MWER alignment
(Matusov et al 2005)

Applied on system outputs En->Fr (Karakanta et al. 2020)
1) NMT, 2) ST cascade, 3) direct ST, 4) pretrained direct ST

## Exp2: What does BLEU$_{br}$ really measure?

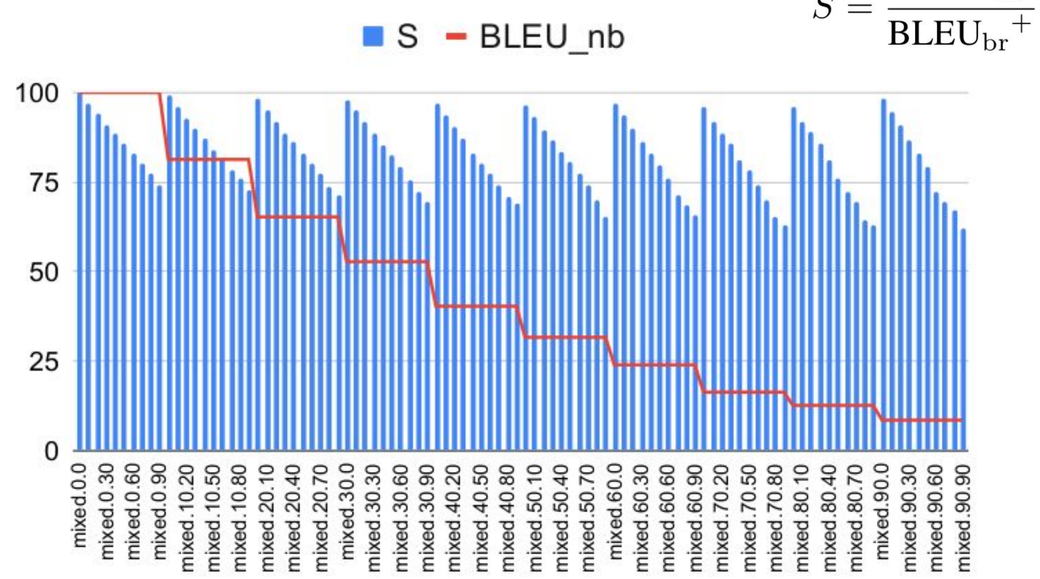
BLEU$_{br}$ largely depends on text quality (BLEU$_{nb}$).
The relative difference with BLEU$_{nb}$ cannot be used to measure segmentation quality in general.




## Sigma: a new segmentation metric

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

$$S = \frac{\mathrm{BLEU_{br}}}{\mathrm{BLEU_{br}}^{+}}$$




System ranking:
NMT>Cas>e2e

| | Projected | | | | | | | | Imperfect | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| System | $P_k$ | Windiff | SegSim | BndSim | Prec | Rec | F1 | $BLEU_{br}$ | $TER_{br}$ | $BLEU_{br}$ | $TER_{br}$ | $S$ |
| NMT | .192 | .208 | .979 | .637 | .711 | .735 | .723 | 83.18 | 6.87 | 32.16 | 19.38 | 89.2 |
| Cas | .252 | .270 | .970 | .519 | .639 | .667 | .653 | 76.14 | 8.91 | 26.34 | 23.23 | 83.1 |
| $e2e_{base}$ | .257 | .277 | .969 | .515 | .601 | .667 | .632 | 75.00 | 9.29 | 22.53 | 24.48 | 81.8 |
| $e2e_{pt}$ | .252 | .276 | .969 | .525 | .610 | .702 | .653 | 74.89 | 9.24 | 26.36 | 23.52 | 81.5 |

Agreement on high quality output

## 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