MUNI
FI

Machine Translation Metrics

Edoardo Signoroni

MUNI
FI

Outline

- Introduction
- Lexical overlap metrics
- Embedding similarity metrics
- Fine-tuned metrics
- Recent developments
- Summary

- How good a machine translation system is?
- **Adequacy**: Does the output convey the same meaning as the input sentence?
- **Fluency**: Is the output good fluent?

Introduction

- Automatic evaluation metrics are commonly used to estimate the quality of a MT system
- They allow for low-cost, fast comparison
- Metrics can be divided in three (four) main types:
  - Lexical overlap metrics
  - Embedding similarity metrics
  - Fine-tuned metrics
  - (Reference-free metrics or Quality Estimation metrics, they compare translation and source without reference. A different task.)

MUNI
FI

Lexical overlap metrics - Intro

- They compare the sequence similarity between the proposed translation and one (or more) reference(s)
- BLEU (Papineni et al. 2002)
- ChrF / ChrF++ (Popovic, 2015, 2017)

MUNI
FI

Lexical overlap metrics - BLEU

- 1- to 4-gram overlap between machine translation output and reference translation
- Computed as precision minus a length penalty for too short translations

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \left(\prod_{i=1}^{4} \textit{precision}_i\right)^{\frac{1}{4}}$$

- Usually computed over the whole corpus, and given as a score between 0 and 100

MUNI
FI

Lexical overlap metrics - BLEU

Hyp1: **A** *whale* **is** *under* **the table**

Hyp2: *The* **cat is** *at* **the table**

Ref: A cat is on the table

- Does not consider word order, nor syntax. It is not suited for morphologically complex languages

MUNI
FI

BLEU is bad

- Routinely scoring among the lowest metrics at WMT Metrics shared task
- Negatively influences the development of MT research
- Increases of 1-2 BLEU do NOT reflect real increase in quality when human judgment is involved
- Nonetheless, BLEU is still used by ~98% of the MT publications (as of 2021)
  - Mathur et al. ACL 2020, Kocmi et al. WMT 2021

Lexical overlap metrics - ChrF

- Averaged F-score over character and word n-grams

$$\mathrm{chrF}_\beta = (1+\beta^2)\frac{\mathrm{chrPchrR}}{\beta^2.\mathrm{chrP}+\mathrm{chrR}}$$

  where β is the weighting of precision and recall.

- Operating at character level helps with morphological variants
- Best lexical overlap metric
- It is advised to use ChrF++(word 2-grams) as a secondary metric for languages unsupported by more advanced metrics

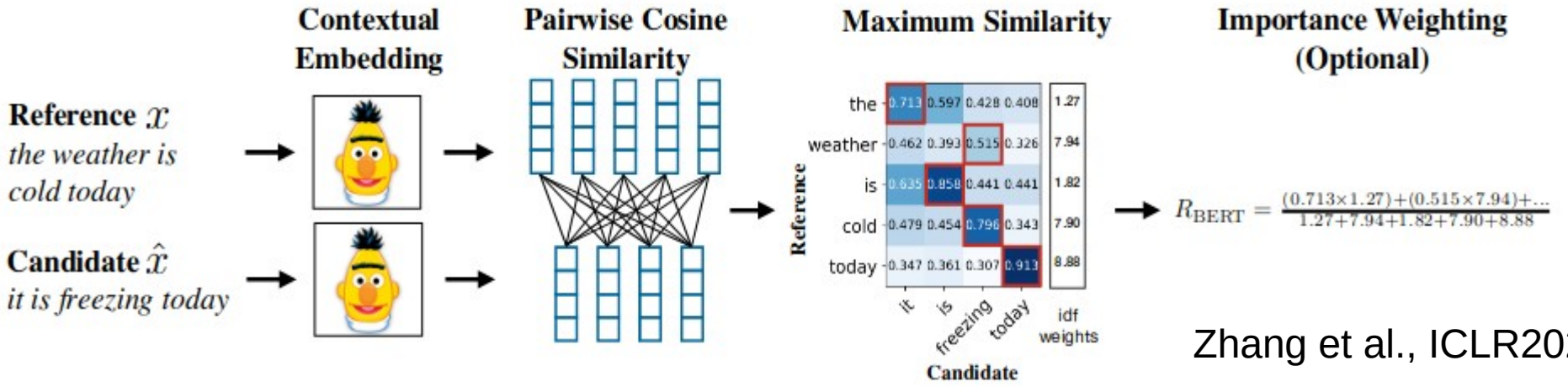
MUNI
FI

Embedding similarity metrics - Intro

- They leverage embedding similarity to account for meaning and compositional diversity, instead of the simple approximations of overlap metrics

Embedding similarity metrics - BERTScore

- Based on pre-trained BERT contextual embeddings
- Creates soft word alignments in candidate and reference with cosine similarity and then returns a precision, recall and F1 score.
- Embeddings are better at capturing distant dependencies, ordering, and allow for soft matching



Zhang et al., ICLR2020

MUNI
FI

Fine-tuned metrics - Intro

- Fine-tuned metrics predict a score based on a given input of source, translation, and reference
- Fine-tune LMs by training on human annotated scores
- The most common frameworks for annotation are:
  - Direct Assessment (DA, Graham et al., 2013) : assign a score between 0 and 100 (or 0 and 1)
  - Multidimensional Quality Metric (MQM, Lommel et al., 2014) : annotate error spans and typology, and then calculate a score. Higher quality, but harder to produce.
- Fine-tuned metrics are the current state-of-the-art

MUNI
FI

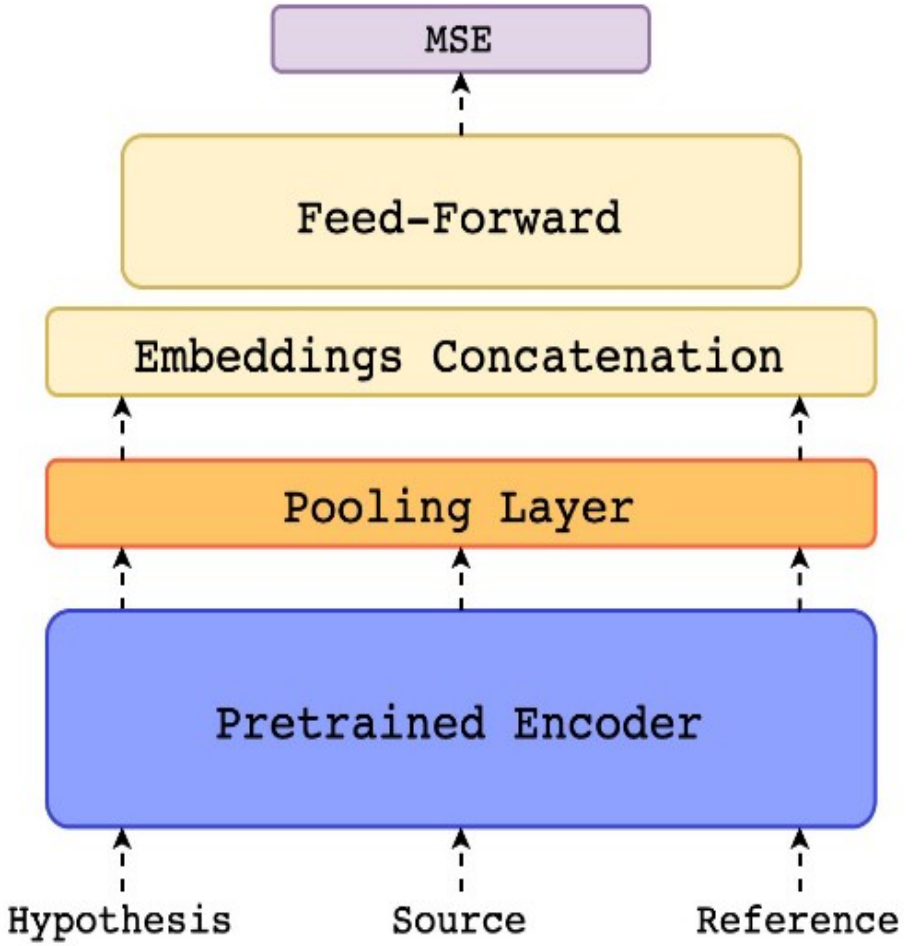
Fine-tuned metrics - BLEURT

- BERT-based metric fine-tuned on DA data
- Pre-trained on a large corpus of synthetic data, e.g. Perturbations of Wikipedia, paraphrases with backtranslation, masked sentences, …
- The pre-training augmented with semantic and lexical-level signals allows the model to generalize better

MUNI
FI

Fine-tuned metrics - COMET

- Fine-tuned XLM-RoBERTa-large on DA data
- Source, translation, and reference are encoded separately, then the output are pooled together
- An estimator layer on top of the encoder outputs the predicted score DA score



Rei et al., EMNLP2020

Recent developments - Intro

Metric	avg rank
MetricX XXL	1.20
COMET-22	1.32
UniTE	1.86
BLEURT-20	1.91
COMET-20	2.36
MATESE	2.57
CometKiwi*	2.70
MS-COMET-22	2.84
UniTE-src*	3.03
YiSi-1	3.27
COMET-QE*	3.33
MATESE-QE*	3.85
MEE4	3.87
BERTScore	3.88
MS-COMET-QE-22*	4.06
chrF	4.70
f101spBLEU	4.97
HWTSC-Teacher-Sim*	5.17
BLEU	5.31
REUSE*	6.69

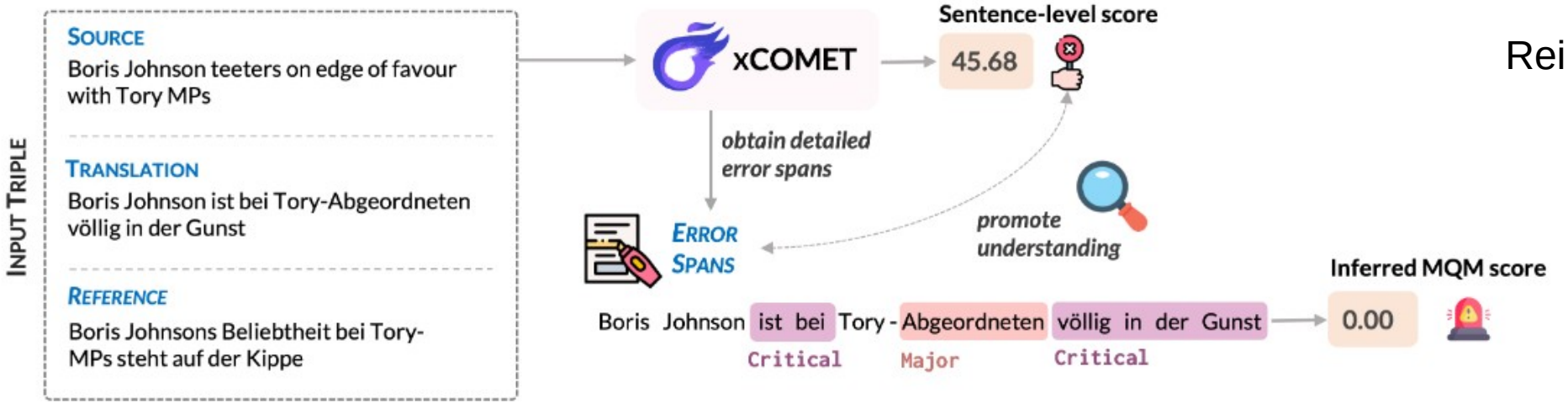
Metric		avg corr
XCOMET-Ensemble	**1**	**0.825**
XCOMET-QE-Ensemble*	2	0.808
MetricX-23	2	0.808
GEMBA-MQM*	2	0.802
MetricX-23-QE*	2	0.800
mbr-metricx-qe*	3	0.788
MaTESe	3	0.782
CometKiwi*	3	0.782
COMET	3	0.779
BLEURT-20	3	0.776
KG-BERTScore*	3	0.774
prismRef	5	0.744
mre-score-labse-regular	5	0.743
BERTscore	5	0.742
XLsim	6	0.719
f200spBLEU	7	0.704
MEE4	7	0.704
tokengram_F	7	0.703
embed_llama	7	0.701
BLEU	7	0.696
chrF	7	0.694
eBLEU	7	0.692
Random-sysname*	8	0.529
prismSrc*	9	0.455

- Results of the WMT 22 & 23
- Fine-tuned metrics score the best
- Recent developments:
  - XCOMET: explainble metric
  - MetricX: encoder-decoder metric

Freitag et al., WMT2022/23

Recent developments - xCOMET

- simultaneously performs sentence-level evaluation and error span detection
- Curriculum training: 1. DA; 2. MQM augmented with synthetic critical errors; 3. high-quality MQM
- State-of-the-art achieved by ensembling 1 XL and 2 XXL checkpoints



Rei et al., WMT2023

MUNI
FI

Recent developments - MetricX-23

- Based on mT5-XXL encoder-decoder
- Fine-tuned on DA and then MQM
- Some interesting insights on metric training:
  - Performance increases with the size of the model
  - Train on z-normalized DA scores and raw scores is a trade off between segment and system level performance
  - Fine-tuning on raw MQM scores is better than z-norm
  - Fine-tuning on DA is better for system-level, fine-tuning on MQM is better for segment-level
- Best metric in WMT22, second-best in WMT23

Juraska et al., WMT2023

Wang et al., 2023 pp

Recent developments – IndicCOMET and AfriCOMET

Sai B et al., ACL2023

- Based on the COMET framework
- **IndicCOMET** fine-tunes COMET-DA on a new MQM dataset for 5 (gu, hi, mr, ml, ta) Indic languages
- **AfriCOMET** builds upon variations of mBERT and XLM-RoBERTa fine-tuned on text and MQM data from typologically diverse African languages
  - They also devise a simplified MQM procedure that can be used by non-experts
- Both model outperform standard COMET for their specific language sets and can zero-shot into related languages

MUNI
FI

Summary

- **Scrap BLEU**
- Use the **newest neural metrics**, such as XCOMET, when possible
- Use ChrF++ as a **secondary** metric, when dealing with unsupported languages
- Devising and training metrics for a specific set of languages is worthwhile and effective

MUNI
FI

NLP Centre

@edo_signoroni

edoardosignoroni.github.io

FACULTAS ARTIS INFORMATICÆ • UNIVERSITAS MASARYKIANA •