MÜ-Qualitätsmetriken der neuen Generation

Vortrag im Rahmen der Fachtagung „Eine Welt ohne Babel“

Fachbereich für Translations-, Sprach- und Kulturwissenschaft (FTSK) Germersheim der Johannes Gutenberg-Universität Mainz

30.06.2022

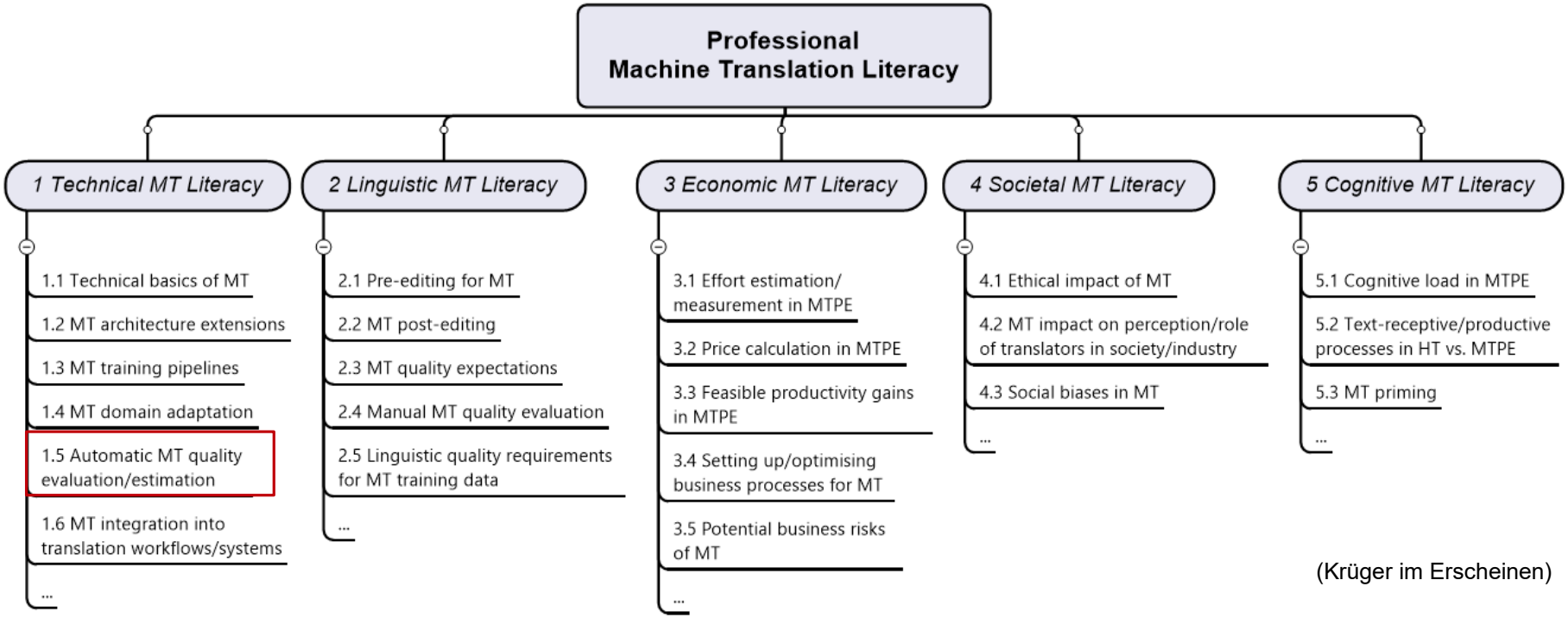
Prof. Dr. Ralph Krüger

Institut für Translation und Mehrsprachige Kommunikation

TH Köln

Technology
Arts Sciences
TH Köln

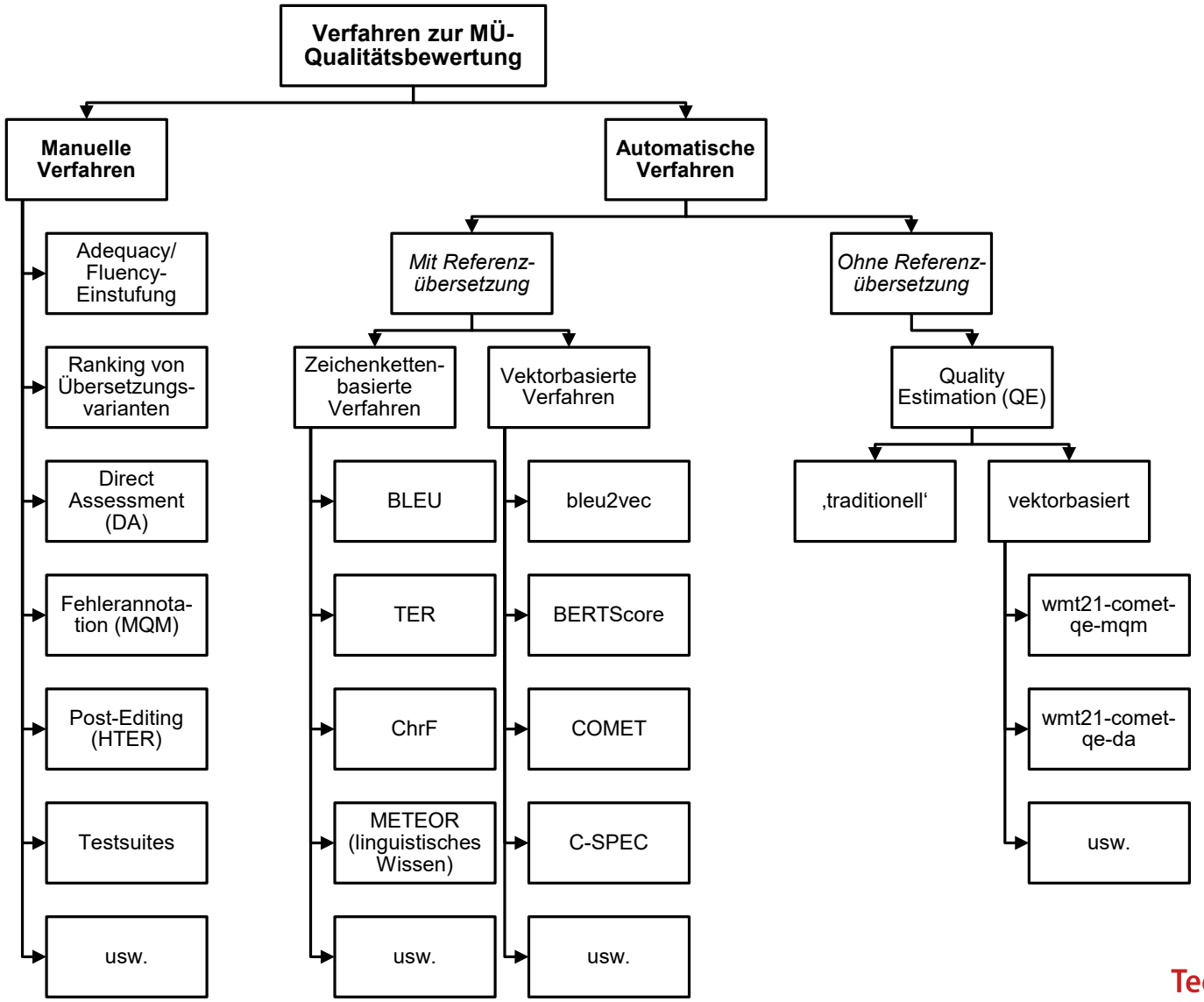
Verortung im Kontext der Professional MT Literacy



(Krüger im Erscheinen)

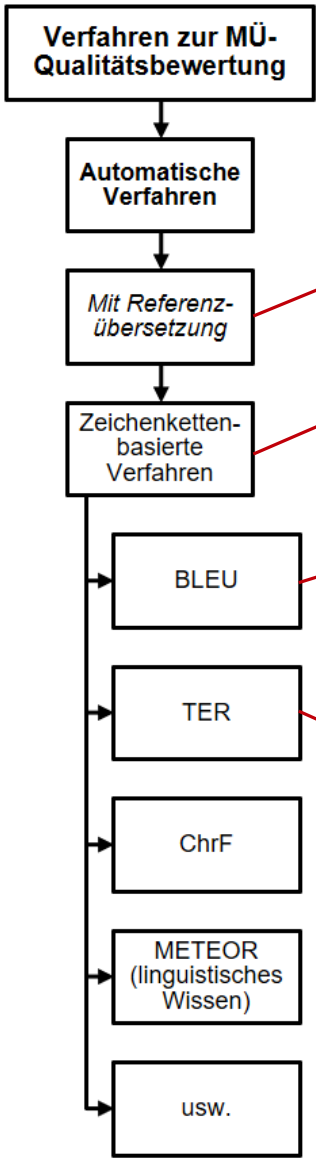
„Knowledge about automatic MT quality *estimation* includes a basic understanding of the principles underlying MT quality estimation as well as its possible applications […]. Knowledge about automatic MT quality *evaluation* covers the basic principles underlying traditional string-matching based proximity measures such as *BLEU* […] or *TER* […] and modern embedding-based metrics such as *COMET* […] as well as **being able to interpret these metrics and to explain what their relative advantages and shortcomings are**." (Krüger im Erscheinen, meine Hervorhebung)

Verfahren zur MÜ-Qualitätsbewertung im Überblick



Technology
Arts Sciences
TH Köln

MÜ-Qualitätsmetriken der ‚alten' Generation



Menschliche Referenzübersetzung = ‚Goldstandard'

Erfasst werden formale (nicht semantische) Ähnlichkeiten (Ausnahme: METEOR)

Ähnlichkeitsmaß auf Basis von N-Grammen (Papineni et al. 2002):

MÜ-Output: How old are you?

Referenzübersetzung: What is your age?

$$BLEU = min(1, \frac{hypothesis\ length}{reference\ length})(\prod_{i=1}^{n} precision_i)^{\frac{1}{n}}$$

BLEU (smoothing method 3): **0,11**

Distanzmaß zur Messung des Post-Editing-Aufwands (Snover et al. 2006):

MÜ-Output: How old are you?

Referenzübersetzung: What is your age?

$$TER = \frac{substitutions + insertions + deletions + shifts}{reference\ length}$$

TER (wortbasiert): **0,8**

Technology
Arts Sciences
TH Köln

MÜ-Qualitätsmetriken der ‚alten' Generation

Vor- und Nachteile

- Vorteile:

- schnell und leicht zu implementieren (gewisse Ausnahme: METEOR)
- gewisse Korrelation mit menschlichen Qualitätsurteilen
- ‚am Markt etabliert'

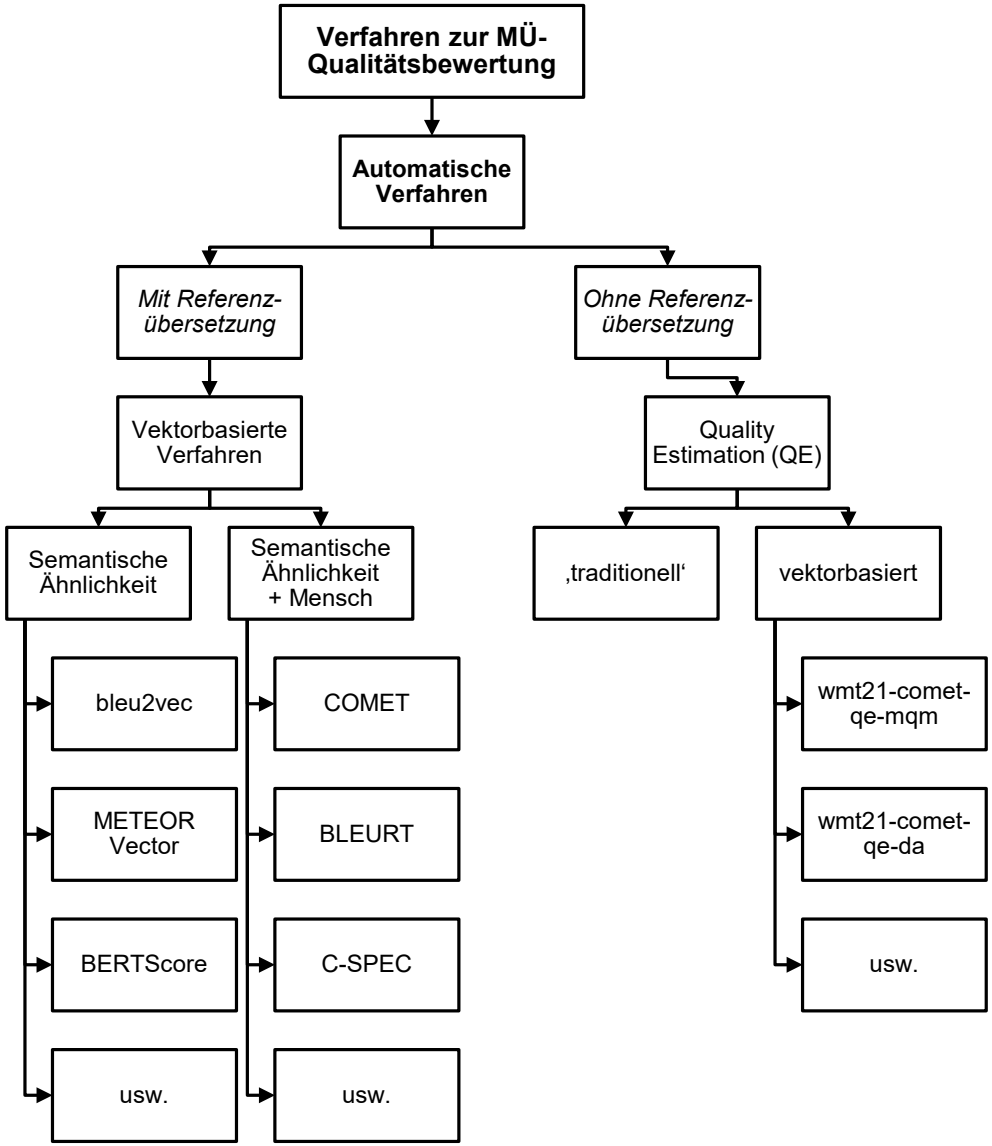
- Nachteile:

- erfasst werden nur formale Ähnlichkeiten zwischen MÜ-Output und Referenz
- Korrelation mit menschlichen Qualitätsurteilen häufig nicht hoch genug
- geringe Qualitätsunterschiede zwischen mehreren hochperformanten Systemen werden oft nicht erfasst
- ‚Mismatch' zwischen Funktionslogik der ‚alten' MÜ-Qualitätsmetriken (Grundbausteine: Zeichenketten) und der Funktionslogik der neuronalen MÜ (Grundbausteine: Wortvektoren)
- Leistungsvorteile neuer KI-basierter NLP-Verfahren werden nicht genutzt

(vgl. Lavie 2021)

MÜ-Qualitätsmetriken der ‚neuen' Generation

Überblick



Technology
Arts Sciences
TH Köln

Vektorbasierte MÜ-Qualitätsmetriken

Grundlagen: Wortvektoren/Word Embeddings

- Word Embeddings = hochdimensionale Vektorrepräsentationen von Wörtern
- Word-Embedding-Modelle werden von speziell dafür vorgesehenen neuronalen Netzen auf Basis großer Textkorpora gelernt

- Beispiel: 100-dimensionale Vektordarstellung des Wortes *translation*:

```
# Display individual word vectors
word_embeddings['translation']
```

```
array([ 0.20131  ,  0.080767 , -0.094986 ,  0.23548  ,  0.33012  ,
        0.26474  ,  0.080543 , -0.7922   ,  0.26269  , -0.043613 ,
       -0.17119  ,  0.18364  ,  0.18348  ,  0.243    , -0.0082833,
        0.56222  ,  0.29671  , -0.2603   ,  0.24834  ,  0.36273  ,
       -0.46913  , -0.6194   ,  0.1258   ,  0.34724  ,  0.3409   ,
       -0.7619   , -0.29654  , -0.15174  ,  0.34333  ,  0.52984  ,
       -0.90529  ,  0.37364  , -0.69871  , -0.42646  , -0.1723   ,
        0.14295  , -0.17003  ,  0.77031  , -0.032785 , -0.24152  ,
       -0.12204  ,  0.36428  , -0.39044  ,  0.49637  , -0.038901 ,
       -0.86653  ,  0.075376 , -0.82226  , -0.23651  ,  0.13407  ,
        0.46013  ,  1.0339   ,  0.65439  ,  0.20841  , -0.41     ,
       -1.2582   , -0.31241  , -0.081561 ,  0.81674  ,  0.12626  ,
        0.3734   ,  0.3857   , -0.31916  , -0.42922  ,  0.61746  ,
       -0.47773  , -0.25266  ,  0.21529  ,  0.40244  , -0.47708  ,
       -0.51625  ,  1.2291   ,  1.1721   , -0.45434  ,  0.19249  ,
        0.11081  , -0.28273  ,  0.77374  , -0.8822   , -0.50809  ,
       -0.27464  , -0.60878  , -0.7705   , -0.0090414, -1.7109   ,
        0.41255  ,  0.021598 , -1.4519   ,  0.42112  , -0.24501  ,
        0.28209  ,  0.35286  ,  0.074104 ,  0.40391  ,  0.32091  ,
       -0.46015  ,  0.19619  , -0.84218  , -0.19712  ,  0.48293  ],
      dtype=float32)
```

(Word-Embedding-Modell: glove-wiki-gigaword-100,
Gensim data-repository: https://github.com/RaRe-Technologies/gensim-data)

Technology
Arts Sciences
TH Köln

Vektorbasierte MÜ-Qualitätsmetriken

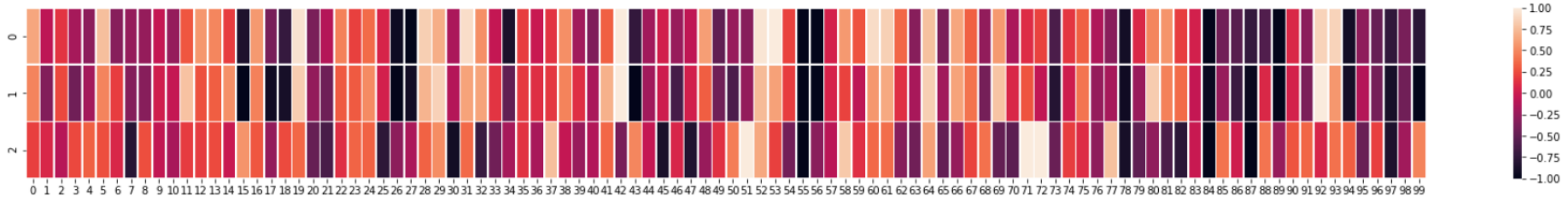
Grundlagen: Wortvektoren/Word Embeddings

- Semantisch ähnliche Wörter haben ähnliche Vektordimensionen

- Beispiel: 100 Vektordimensionen *father/son* vs. 100 Vektordimensionen *translation*:

```
# Import the libraries required to visualise the word embedding vectors
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

# Create the visual representations
plt.figure(figsize=(30,3))
sns.heatmap([word_embeddings["father"], word_embeddings["son"], word_embeddings["translation"]], xticklabels=True, yticklabels=True, cbar=True,
            vmin=-1, vmax=1, linewidths=0.7)
plt.show()
```



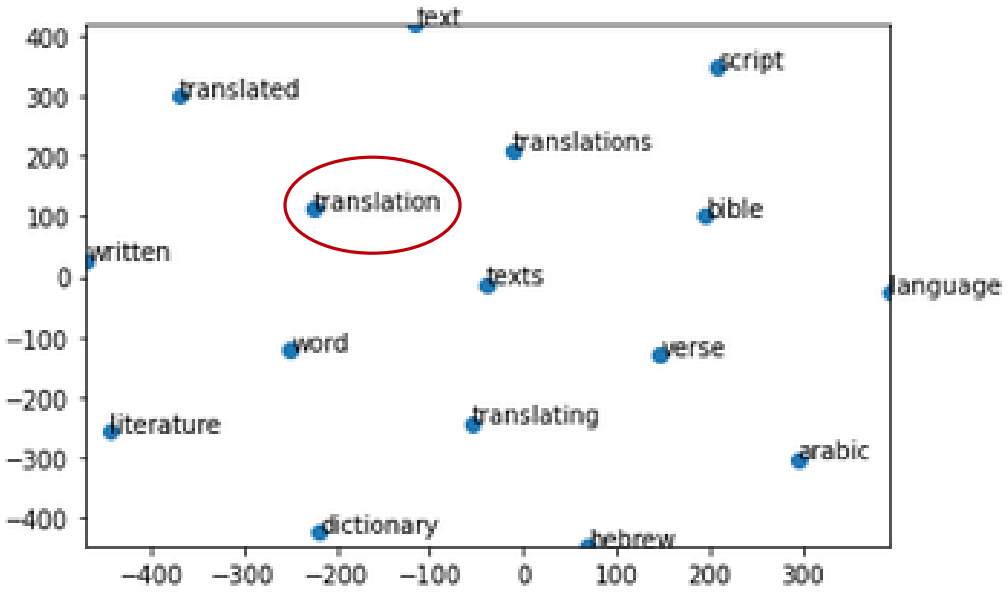
Technology
Arts Sciences
TH Köln

Vektorbasierte MÜ-Qualitätsmetriken

Grundlagen: Wortvektoren/Word Embeddings

- Semantische Ähnlichkeit von Wörtern = räumliche Nähe der entsprechenden Wortvektoren im Vektorraum

- Beispiel: 15 Wörter mit größter semantischer Ähnlichkeit/räumlicher Nähe zum Wort *translation* in zweidimensionalem Vektorraum:



Technology
Arts Sciences
TH Köln

Vektorbasierte MÜ-Qualitätsmetriken

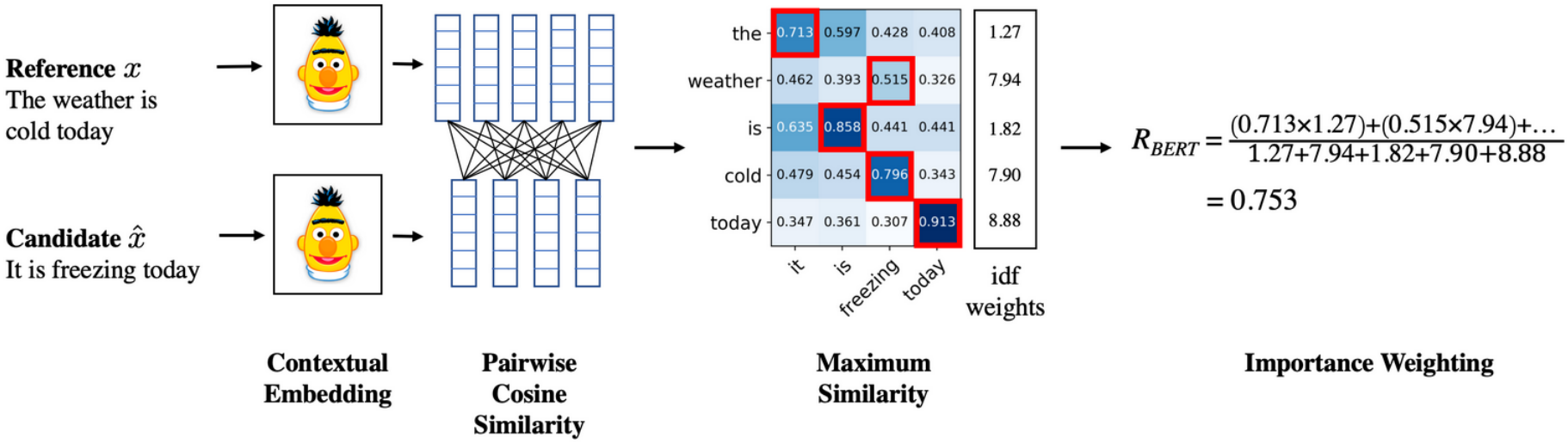
Semantische Ähnlichkeit: *BERTScore* (Zhang et al. 2020)

- Grundlage: Google BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers, Devlin et al. 2019)

→ neuronales Sprachmodell auf Basis der Transformer-Architektur für NMÜ-Systeme (Vaswani et al. 2017)

- Berechnung von Precision, Recall u. F-Measure auf Grundlage kontextualisierter Word Embeddings

- BERTScore basiert auf einsprachigem Vektorraum



(Zhang et al. 2020:4)

Vektorbasierte MÜ-Qualitätsmetriken

Semantische Ähnlichkeit: *BERTScore* (Zhang et al. 2020)

- Beispiel für Erfassung semantischer Ähnlichkeit bei formaler Unterschiedlichkeit durch BERTScore:

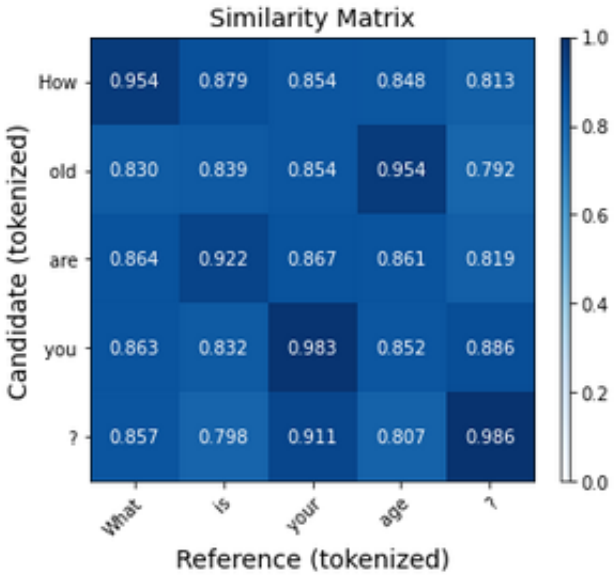
MÜ-Output: How old are you?

Referenz: What is your age?

BLEU (Ähnlichkeitsmaß): 0,11

TER (Distanzmaß): 0,8

BERTScore (Ähnlichkeitsmaß): **0,761**



Technology
Arts Sciences
TH Köln

Vektorbasierte MÜ-Qualitätsmetriken

Von reiner semantischer Ähnlichkeit zu semantischer Ähnlichkeit + menschliche Qualitätsurteile

„Embedding-based metrics like METEOR-VECTOR […], BLEU2VEC […], YISI-1 […], MOVERSCORE […], and BERTSCORE […] create soft-alignments between reference and hypothesis in an embedding space and then compute a score that reflects the semantic similarity between those segments. However, **human judgements such as DA and MQM, capture much more than just semantic similarity**, resulting in a correlation upper-bound between human judgements and the scores produced by such metrics.“
(Rei et al. 2020:2692, meine Hervorhebung)

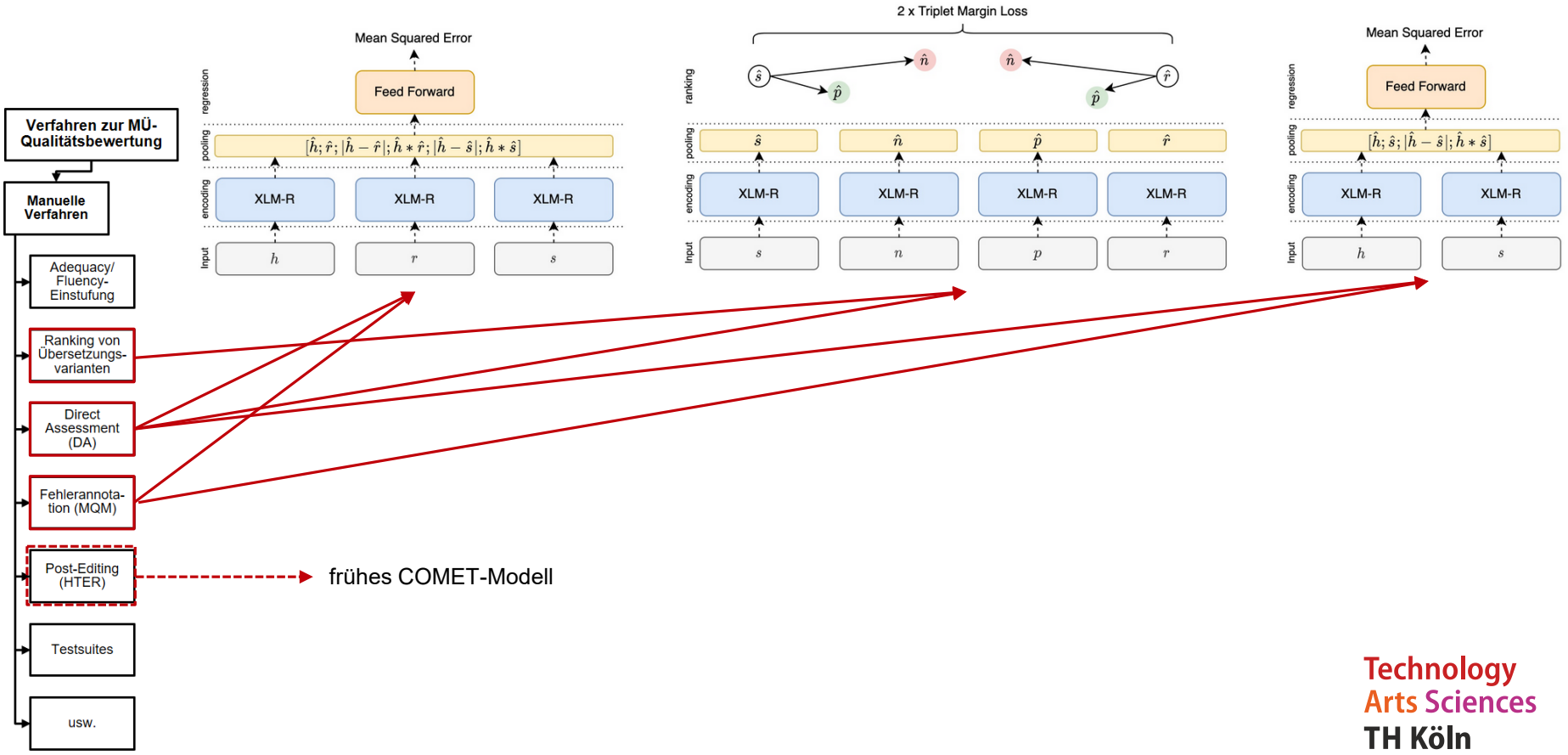
Vektorbasierte MÜ-Qualitätsmetriken

semantische Ähnlichkeit + menschliche Qualitätsurteile: *COMET* (Rei et al. 2020)

- **C**rosslingual **O**ptimized **M**etric for **E**valuation of **T**ranslation: basiert auf XLM-RoBERTa (Conneau et al. 2020) → mehrsprachiger Vektorraum mit mehrsprachigen Embeddings

- COMET-Architekturen und Quellen für menschliche Qualitätsurteile:

(https://unbabel.github.io/COMET/html/models.html)



Technology
Arts Sciences
TH Köln

Vektorbasierte MÜ-Qualitätsmetriken

COMET-Modelle bei der WMT 2021 Metrics Shared Task

- Ergebnisse der WMT 2021 Metrics Shared Task für ZH-EN und EN-DE auf Segmentebene:

	N° Segments	zh-en 4400		en-de 2950			
		Pearson	Kendall	Pearson	Kendall	Pearson Avg.	Kendall Avg.
Baselines	BLEURT	0.492	0.405	0.107	0.060	0.299	0.232
	PRISM	0.399	0.337	0.072	0.020	0.235	0.178
	BERTSCORE	0.441	0.344	0.116	0.060	0.279	0.202
	BLEU	0.196	0.275	0.062	0.004	0.129	0.140
	CHRF	0.267	0.219	0.119	0.059	0.193	0.139
	COMET-DA (2020)	0.538	0.435	0.425	0.282	0.481	0.359
Ref. based	COMET-DA (2021)	0.559	0.454	0.464	0.309	0.511	0.382
	COMET-MQM (2021)	**0.717**	**0.546**	**0.488**	**0.361**	**0.602**	**0.454**
	COMETINHO-DA	0.484	0.386	0.299	0.204	0.392	0.295
	COMETINHO-MQM	0.670	0.496	0.311	0.237	0.490	0.367
Ref. Free	COMET-QE-DA (2021)	0.567	0.436	**0.497**	0.308	0.532	0.372
	COMET-QE-MQM (2021)	**0.720**	**0.531**	0.470	**0.359**	**0.595**	**0.445**
	OPENKIWI	0.522	0.385	0.448	0.287	0.485	0.336

Table 2: Segment-level correlations on the *en-de* and *zh-en* testset.

(Rei et al. 2021:1033)

Vektorbasierte MÜ-Qualitätsmetriken

Ad-hoc-Test von COMET-QE-MQM mit fachübersetzungsrelevanten Phänomenen EN-DE

- Sekundäre Subjektivierung

AT: *This report tells the story of plastics*, and their effect on the environment and climate.

DeepL: *Dieser Bericht erzählt die Geschichte der Kunststoffe* und ihrer Auswirkungen auf die Umwelt und das Klima.

HÜ: *In diesem Bericht wird die Geschichte der Kunststoffe* und ihrer Auswirkungen auf die Umwelt und das Klima *erzählt*.

COMET-QE-MQM DeepL: 0,14950

COMET-QE-MQM HÜ: **0,14996**

- Funktionale Satzperspektive

AT: Choose the Settings command from the Tools menu in order to configure the program settings.

DeepL: Wählen Sie den Befehl Einstellungen aus dem Menü Extras, um die Programmeinstellungen zu konfigurieren.

HÜ: Wenn Sie die Programmeinstellungen konfigurieren möchten, klicken Sie im Menü Extras auf den Befehl Einstellungen.

COMET-QE-MQM DeepL: 0,13143

COMET-QE-MQM HÜ: **0,13383**

Vektorbasierte MÜ-Qualitätsmetriken

Ad-hoc-Test von COMET-QE-MQM mit fachübersetzungsrelevanten Phänomenen EN-DE

- Propositional Opaqueness/Explikation

AT: The software can download several files simultaneously and *resume unfinished files*.

DeepL: Die Software kann mehrere Dateien gleichzeitig herunterladen und *nicht abgeschlossene Dateien wieder aufnehmen*.

HÜ: Die Software kann mehrere Dateien gleichzeitig herunterladen und *den Download von noch nicht vollständig heruntergeladenen Dateien wieder aufnehmen*.

COMET-QE-MQM DeepL: 0,12626

COMET-QE-MQM HÜ: **0,13609**

- Dynamische Modalität

AT: Materials *can be classified* according to their electrical behavior as follows:

DeepL: Die Materialien *können* nach ihrem elektrischen Verhalten wie folgt *klassifiziert werden*:

HÜ: Die Materialien *werden* nach ihrem elektrischen Verhalten wie folgt *klassifiziert*:

COMET-QE-MQM DeepL: 0,17456

COMET-QE-MQM Human: **0,17513**

MÜ-Qualitätsmetriken der ‚neuen' Generation

Vor- und Nachteile

- Vorteile:

- Mehrere Studien zeigen höhere Korrelation mit menschlichen Qualitätsurteilen als zeichenkettenbasierte Metriken (TOP3 laut WMT21: C-SPEC$_{PN}$, BLEURT-20 und COMET-MQM_2021, Freitag et al. 2021)
- Auch Erfassung geringer Qualitätsunterschiede zwischen mehreren leistungsstarken Systemen
- Training mit eigenen Datenbeständen möglich (vgl. domänenadaptierte MÜ)
- Ad-hoc-Test legt nahe, dass COMET auch fachübersetzungsrelevante Phänomene in bestimmten Sprachkombinationen erkennen kann

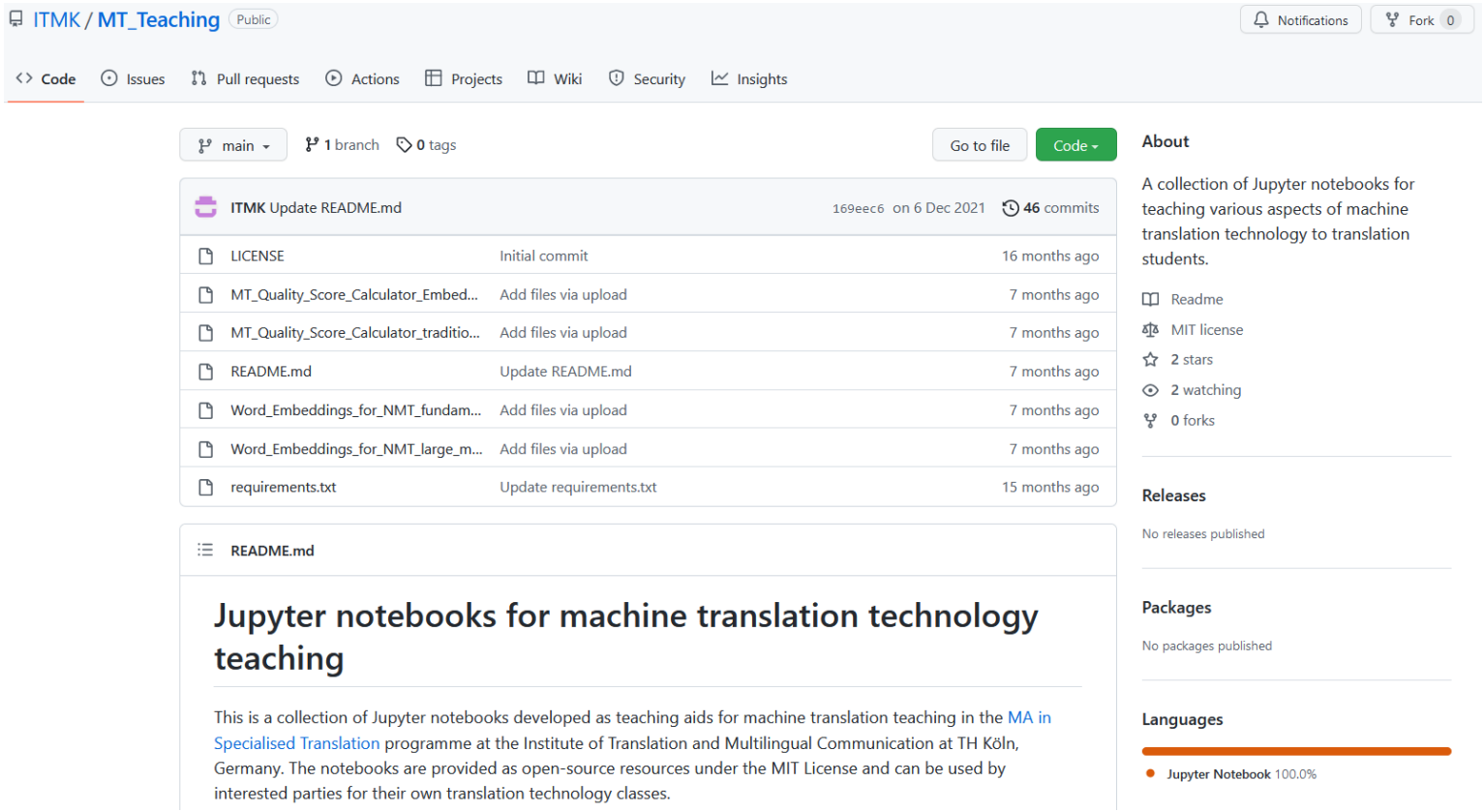
- Nachteile:

- Vektorbasierte Verfahren arbeiten mit großen Sprachmodellen → große Datenvolumina, lange Ladezeiten (aber: *COMETinho*, Rei et al. 2021)
- Implementierung mitunter komplizierter als bei BLEU, TER usw. (insb. COMET)
- Interpretation der Scores teilweise komplizierter als bei BLEU, TER usw.
- Weniger transparent als zeichenkettenbasierte Verfahren
- COMET: aktuell noch Schwächen bei Zahlen und Eigennamen (vgl. Amrhein/Sennrich 2022)
- insgesamt: derzeit noch Schwächen bei Negation und *Sentiment Polarity* (vgl. Freitag et al. 2021)

Anhang

GitHub-Repository mit Jupyter Notebooks zur Translationstechnologielehre

- Popularisierende Darstellung von (alten u. neuen) MÜ-Qualitätsmetriken und Word Embeddings für Studierende (und Lehrende) der Translationswissenschaft (Krüger 2021a, 2021b):



(https://github.com/ITMK/MT_Teaching)

Technology
Arts Sciences
TH Köln

Literatur

- Amrhein, Chantal; Sennrich, Rico (2022): Identifying Weaknesses in Machine Translation Metrics through Minimum Bayes Risk Decoding: A Case Study for COMET: *arxiv*. https://arxiv.org/abs/2202.05148
- Conneau, Alexis; Khandelwal, Kartikay; Goyal, Naman; Chaudhary, Vishrav; Wenzek, Guillaume; Guzmán, Francisco; Grave, Edouard; Ott, Myle; Zettlemoyer, Luke; Stoyanov; Veselin (2020): Unsupervised Cross-lingual Representation Learning at Scale. In: Jurafsky, Dan; Chai, Joyce; Schluter, Natalie; Tetreault, Joel (Hrsg.): *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*, 8440-8451. https://aclanthology.org/2020.acl-main.747/
- Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (2019): BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In: Burstein, Jill; Doran, Christy; Solorio, Thamar (Hrsg.): *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171-4186. https://aclanthology.org/N19-1423/
- Freitag, Markus; Rei, Ricardo; Mathur, Nitika; Lo, Chi-kiu; Stewart, Craig; Foster, George; Lavie, Alon; Bojar, Ondrej: Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain. In: Barrault, Loic; Bojar, Ondrej; Bougares, Fethi; Chatterjee, Rajen; Costa-jussa, Marta R.; Federmann, Christian; Fishel, Mark; Fraser, Alexander; Freitag, Markus; Graham, Yvette; Grundkiewicz, Roman; Guzman, Paco; Haddow, Barry; Huck, Matthias; Jimeno Yepes, Antonio; Koehn, Philipp; Kocmi, Tom; Martins, Andre; Morishita, Makoto; Monz Christof (Hrsg.): *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, 733-774. https://aclanthology.org/2021.wmt-1.73/

Literatur

- Krüger, Ralph (2021a): An Online Repository of Python Resources for Teaching Machine Translation to Translation Students. *Current Trends in Translation Teaching and Learning E* 2021, 4-30. https://doi.org/10.51287/cttle20212
- Krüger, Ralph (2021b): Using Jupyter Notebooks as Didactic Instruments in Translation Technology Teaching. *The Interpreter and Translator Trainer*, 1-21. https://doi.org/10.1080/1750399X.2021.2004009
- Krüger, Ralph (im Erscheinen): Integrating Professional Machine Translation Literacy and Data Literacy. *Lebende Sprachen*.
- Lavie, Alon (2021): COMET: A Neural Framework for State-of-the-Art MT Evaluation. *LTI Colloquium 2020-21*. Carnegie Mellon University. Language Technologies Institute. https://www.youtube.com/watch?v=FrOmIZSStvk
- Papineni, Kishore; Roukos, Salim; Ward, Todd; Zhu; Wie-Jing (2002): BLEU: A Method for Automatic Evaluation of Machine Translation. In: Isabelle, Pierre; Charniak, Eugene; Lin, Dekang (Hrsg.): *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 311-318. https://aclanthology.org/P02-1040/
- Rei, Ricardo; Stewart, Craig; Farinha, Ana C.; Lavie, Alon (2020): COMET: A Neural Framework for MT Evaluation. In: Webber, Bonnie; Cohn, Trevor; He, Yulan; Liu, Yang (Hrsg): *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2685-2702. https://aclanthology.org/2020.emnlp-main.213/

Literatur

- Rei, Ricardo; Farinha, Ana C.; Zerva, Chrysoula; van Stigt, Daan; Stewart, Craig; Ramos, Pedro; Glushkova, Taisiya; Martins, André F. T.; Lavie, Alon (2021): Are References Really Needed? Unbabel-IST 2021 Submission for the Metrics Shared Task. In: Barrault, Loic; Bojar, Ondrej; Bougares, Fethi; Chatterjee, Rajen; Costa-jussa, Marta R.; Federmann, Christian; Fishel, Mark; Fraser, Alexander; Freitag, Markus; Graham, Yvette; Grundkiewicz, Roman; Guzman, Paco; Haddow, Barry; Huck, Matthias; Jimeno Yepes, Antonio; Koehn, Philipp; Kocmi, Tom; Martins, Andre; Morishita, Makoto; Monz Christof (Hrsg.): *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, 1030-1040. https://aclanthology.org/2021.wmt-1.111/
- Snover, Matthew; Dorr, Bonnie; Schwartz, Rich; Micciulla, Linnea; Makhoul, John (2006): A Study of Translation Edit Rate with Targeted Human Annotation. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. Association for Machine Translation in the Americas, 223-231. https://aclanthology.org/2006.amta-papers.25/
- Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jacob; Jones, Llion; Gomez, Aidan N.; Kaiser, Łukasz; Polosukhin, Illia (2017): Attention Is All You Need. In: Guyon, Isabelle; Luxburg, Ulrike von; Bengio, Samy; Wallach, Hanna M.; Fergus, Rob; Vishwanathan, S. V. N.; Garnett, Roman (Hrsg.): *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 1-11. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
- Zhang, Tianyi; Kishore, Varsha; Wu, Felix; Weinberger, Kilian Q; Artzi, Yoav (2020): BERTScore: Evaluating Text Generation with BERT. *arXiv*. https://arxiv.org/abs/1904.09675