

Embarrassingly Easy Document-Level MT Metrics: How to Convert Any Pretrained Metric Into a Document-Level Metric

IST & Unbabel seminar

EPFL



Amazon AI

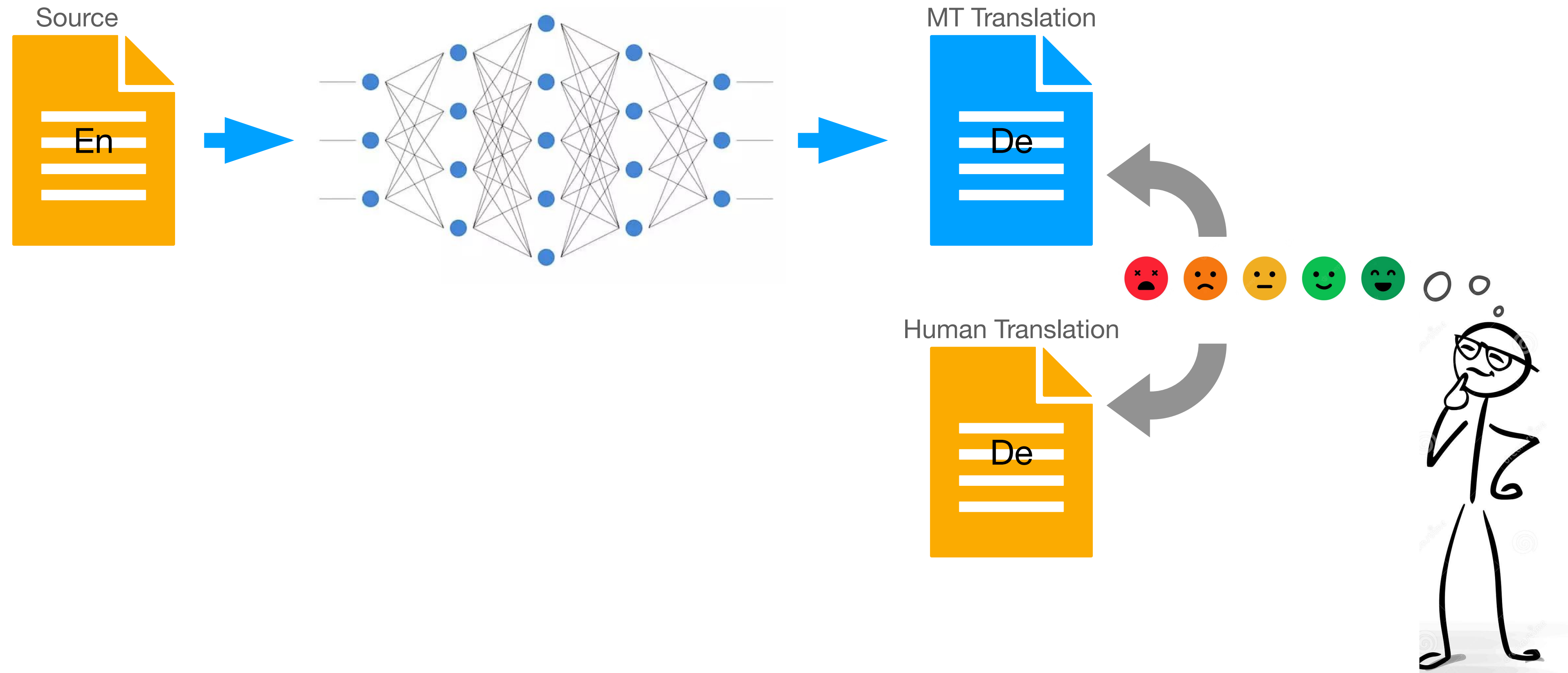
Giorgos Vernikos

Brian Thompson

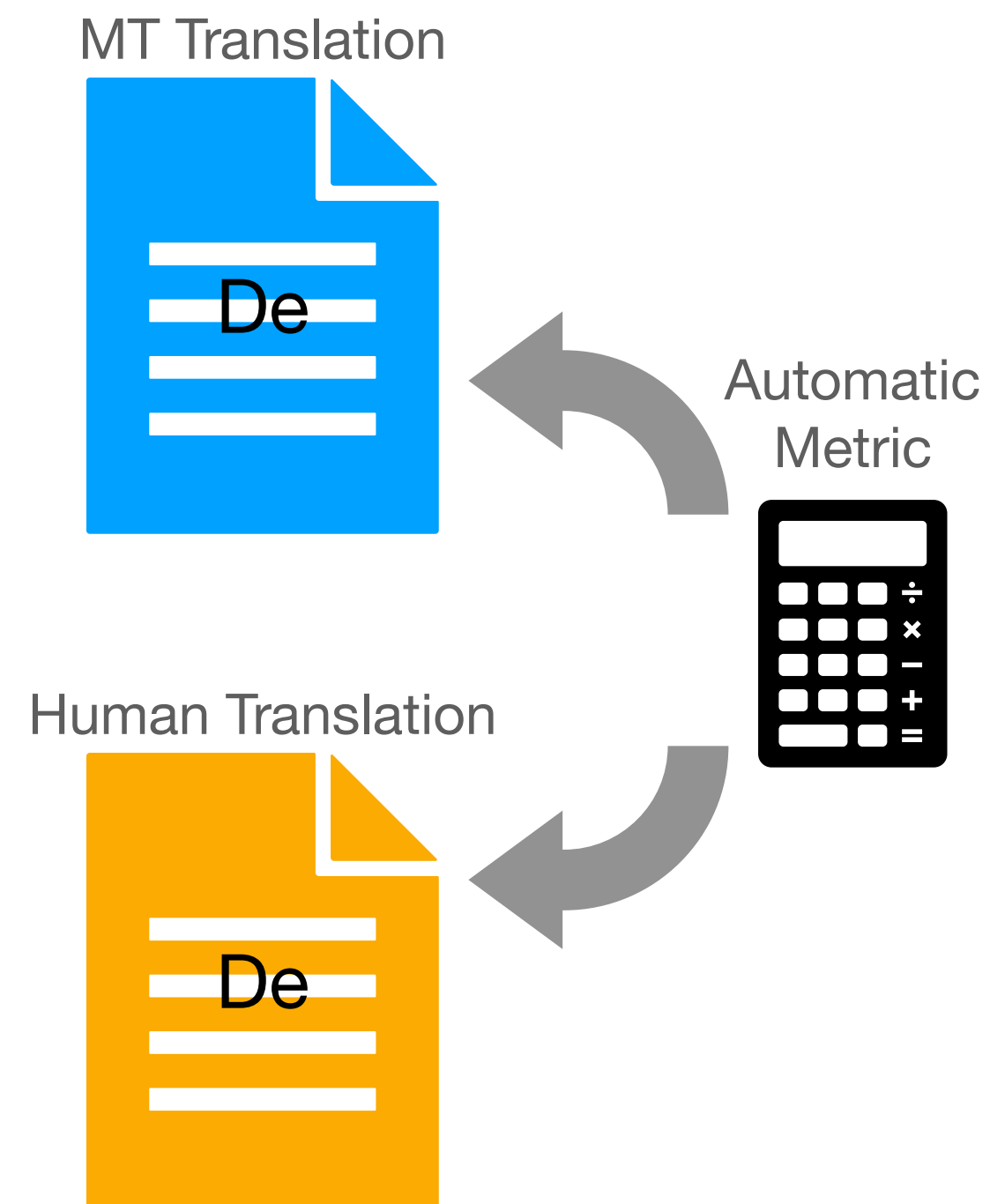
Prashant Mathur

Marcello Federico

Evaluation of Machine Translation



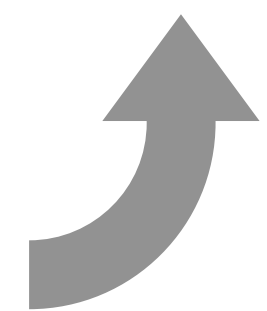
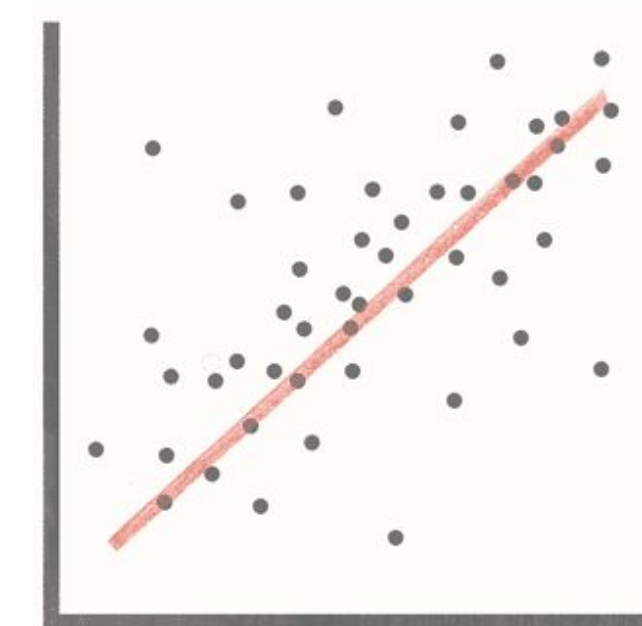
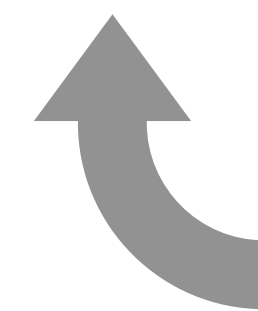
Evaluation of Machine Translation



- Classic, n-gram matching metrics: BLEU^[1], ChrF^[2], TER^[3]
- Recent, learnable metrics: BERTScore^[4], COMET^[5], Prism^[6]

Evaluation of Evaluation of Machine Translation

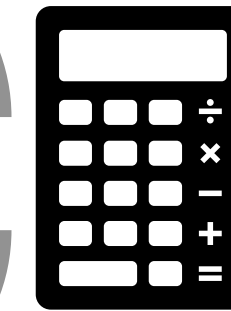
	Human Score (↑)	Metric Score (↑)
Brand in französischem Chemiewerk gelöscht		
Fire extinguished in French chemical plant		
Fire extinguished at French chemical plant	0.57	37.99
Fire at French chemical plant extinguished	0.65	53.73
Brand in French chemistry	-1.83	19.38



Evaluation of Evaluation of Machine Translation

Fire extinguished in French chemical plant

Brand in French chemistry



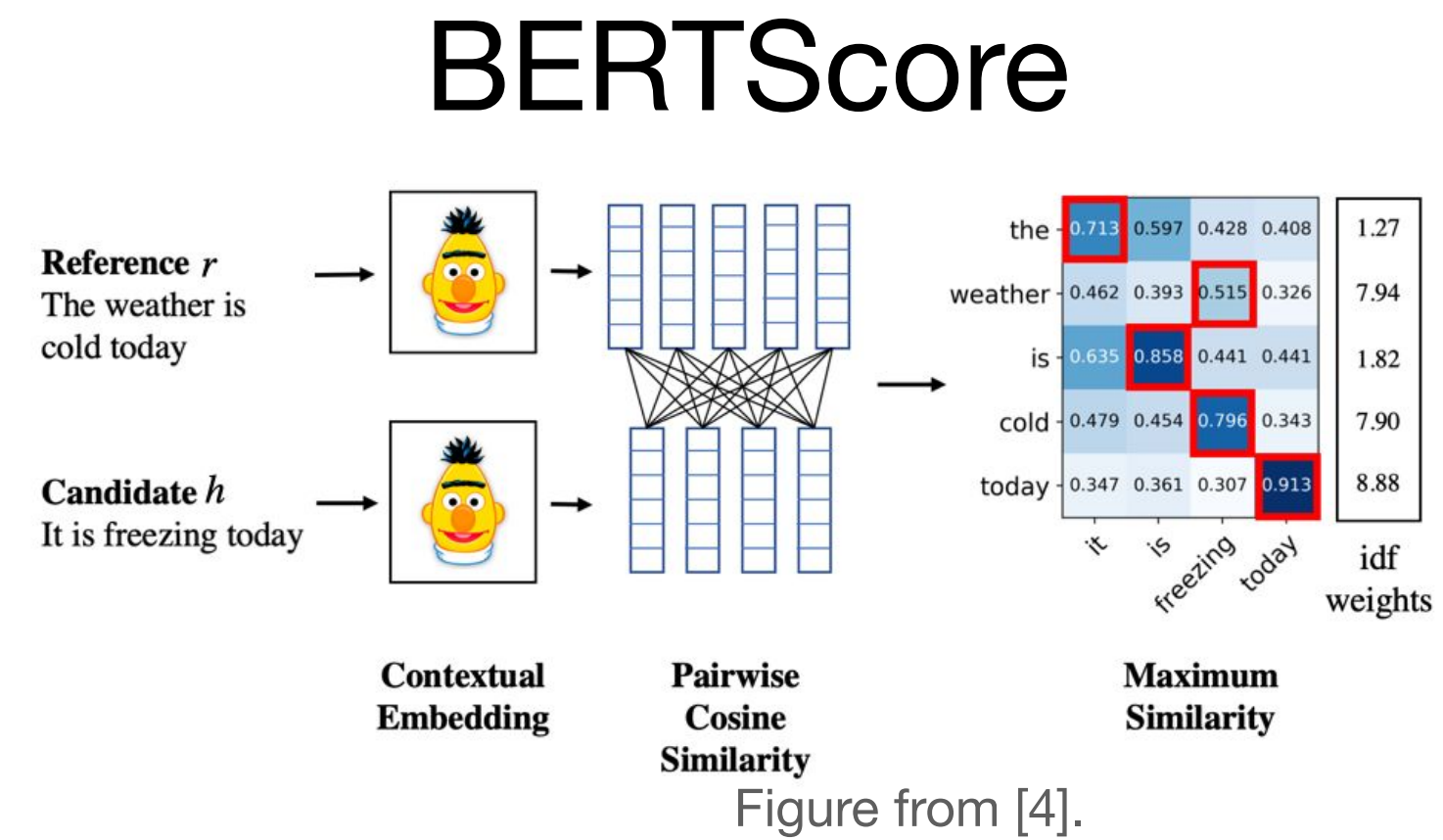
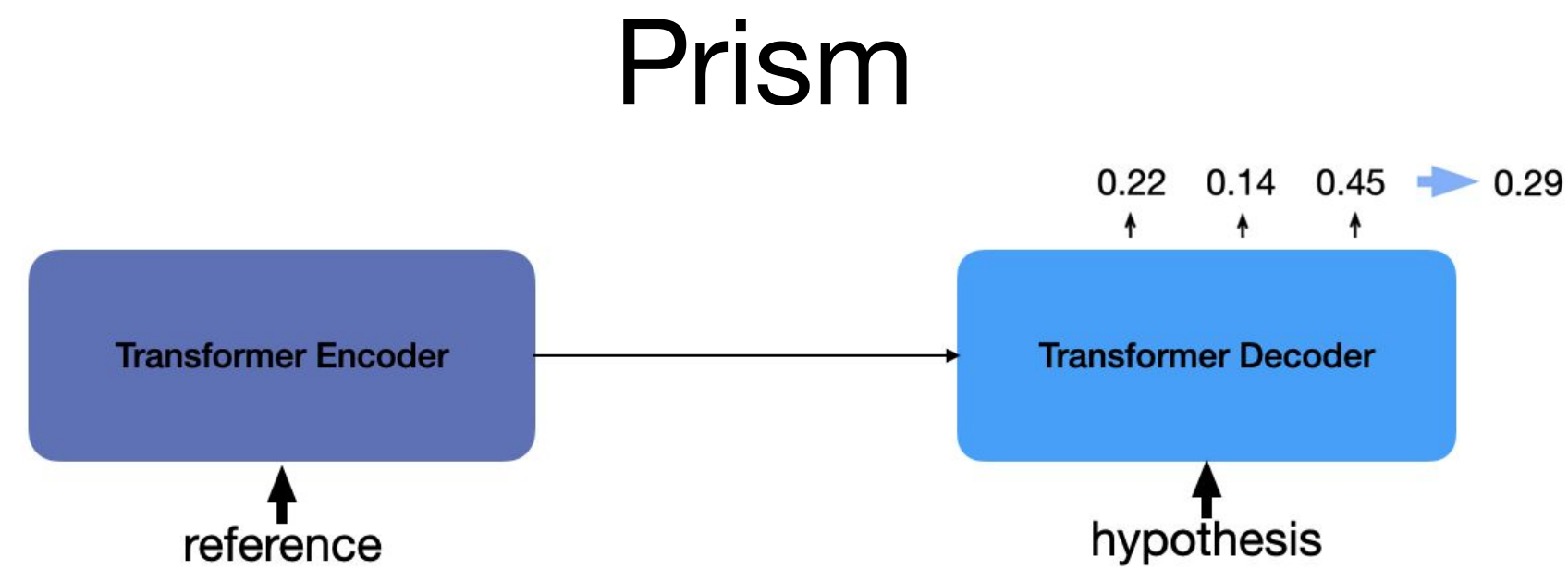
BLEU
ChrF
TER

Why do we need all these metrics ???

Traditional metrics like BLEU demonstrate **poor correlation** with human judgements that can even be **negative** when looking at the top k systems^[7].

Evaluation of Evaluation of Machine Translation

State-of-the-art metrics use representations from pretrained Language Models or MT systems to evaluate MT outputs



COMET(-QE)

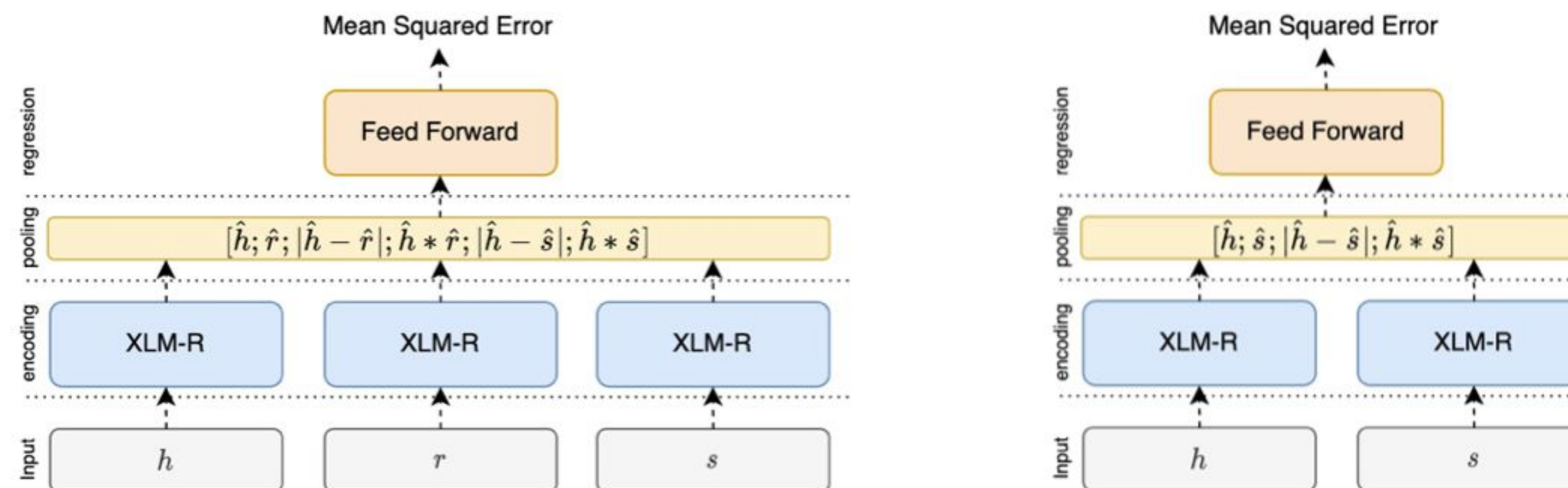
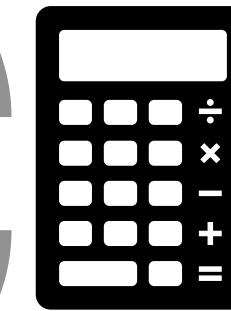


Figure from [5].

Evaluation of Evaluation of Machine Translation

Fire extinguished in French chemical plant

Brand in French chemistry



Prism
BERTScore
COMET

Why do we need all these metrics ???

Metrics that use contextual representations from neural networks have been shown to **correlate better** with humans^[7]!

What is still missing ???

Problem Formulation

Sentences can be ambiguous when judged in isolation !

source-based evaluation

pronoun translation

sent1	I put it in my car.	what is "it"?
+1 pr.	What did you do with the suitcase?	it=SUITCASE
	I put it in my car.	

Figure from [8].

disambiguation

sent2	Yes, she did .	main verb?
+1 pr.	Did she give you any ? Yes, she did .	main verb=GIVE what is "any"?
+2 pr.	So you went to your wife for money . Did she give you any ? Yes, she did .	main verb=GIVE any=MONEY

Figure from [8].

Problem Formulation

Sentences can be ambiguous when judged in isolation !

source-based evaluation

pronoun translation

sent1	I put it in my car.	what is "it"?
+1 pr.	What did you do with the suitcase ?	it=SUITCASE
	I put it in my car.	

Figure from [8].

disambiguation

sent2	Yes, she did .	main verb?
+1 pr.	Did she give you any ? Yes, she did .	main verb=GIVE what is "any"?
+2 pr.	So you went to your wife for money . Did she give you any ? Yes, she did .	main verb=GIVE any=MONEY

Figure from [8].

reference-based evaluation

~~pronoun translation~~

system	translation
human	There are too many bugs .
system1	There are too many insects .
system2	There are too many flaws .
system3	There are too many hidden microphones .

disambiguation

Problem Formulation

Sentences can be ambiguous when judged in isolation !

source-based evaluation

pronoun translation

sent1	I put it in my car.	what is "it"?
+1 pr.	What did you do with the suitcase ?	it=SUITCASE
	I put it in my car.	

Figure from [8].

disambiguation

sent2	Yes, she did .	main verb?
+1 pr.	Did she give you any ? Yes, she did .	main verb=GIVE what is "any"?
+2 pr.	So you went to your wife for money . Did she give you any ? Yes, she did .	main verb=GIVE any=MONEY

Figure from [8].

reference-based evaluation

~~pronoun translation~~

disambiguation

system	translation		
human	There are too many bugs .	+1 pr	Do you ever clean this house?
system1	There are too many insects .	✓	
system2	There are too many flaws .	✗	
system3	There are too many hidden microphones .	✗	

Problem Formulation

Evaluating at the sentence level is **misleading**: MT systems appear to perform better and even reach human parity^[10]

Best practices for human evaluation of MT have been revised and now annotators are strongly advised to take context into account^[11]!

1/12 documents, 4 items left in document WMT20DocSrcDA #214: Doc. #seattle_times.7674-2 English → German (deutsch)

Below you see a document with 6 sentences in English and their corresponding candidate translations in German (deutsch). Score each candidate translation in the document context, answering the question:

How accurately does the candidate text (right column, in bold) convey the original semantics of the source text (left column) in the document context?

You may revisit already scored sentences and update their scores at any time by clicking at a source text.

Expand all items Expand unannotated Collaps all items

Man gets prison after woman finds bullet in her skull	Der Mann wird gefangen, nachdem die Frau in ihrem Schädel geschossen ist	✓
A Georgia man has been sentenced to 25 years in prison for shooting his girlfriend, who didn't realize she survived a bullet to the brain until she went to the hospital for treatment of headaches.	Ein georgischer Mann wurde zu 25 Jahren Gefängnis verurteilt, weil er seinen Freund geschossen hat, der nicht gewusst hatte, dass er eine Kugel ins Gehirn überlebte, bis er in das Krankenhaus zur Behandlung	✓
News outlets report 39-year-old Jerrontae Cain was sentenced Thursday on charges including being a felon in possession of a gun in the 2017 attack on 42-year-old Nicole Gordon.	Nachrichtenagenturen-Bericht 39-jährige Jerrontae Cain wurde am Donnerstag wegen Anklage verurteilt, darunter ein Felon im Besitz einer Waffe beim Angriff auf 42-jährige Nicole Gordon im Jahr 2017.	
Suffering from severe headaches and memory loss, Gordon was examined last year by doctors who found a bullet lodged in her skull.	Gordon, das an schweren Kopfschmerzen und Gedächtnisverlusten leidet, wurde im vergangenen Jahr von Ärzten untersucht, die ein in ihren Schädel eingesetztes Geschoss gefunden haben.	
Gordon told police she didn't remember being shot, but did remember an argument with Cain during which her car window shattered and she passed out. She thought she was hurt by broken glass, and she was patched up at the home of Cain's mother.	Gordon teilte der Polizei mit, dass sie sich nicht daran erinnere, geschossen zu werden, sondern sich an ein Argument mit Cain erinnerte, in dem ihr Autofenster erschütterte und sie ausging. Sie dachte, sie sei von zerbrochenem Glas verletzt worden, und sie wurde in der Heimat der Mutter von Cain aufgesteckt.	

Please score the document translation above answering the question (you can score the entire document only after scoring all previous sentences):

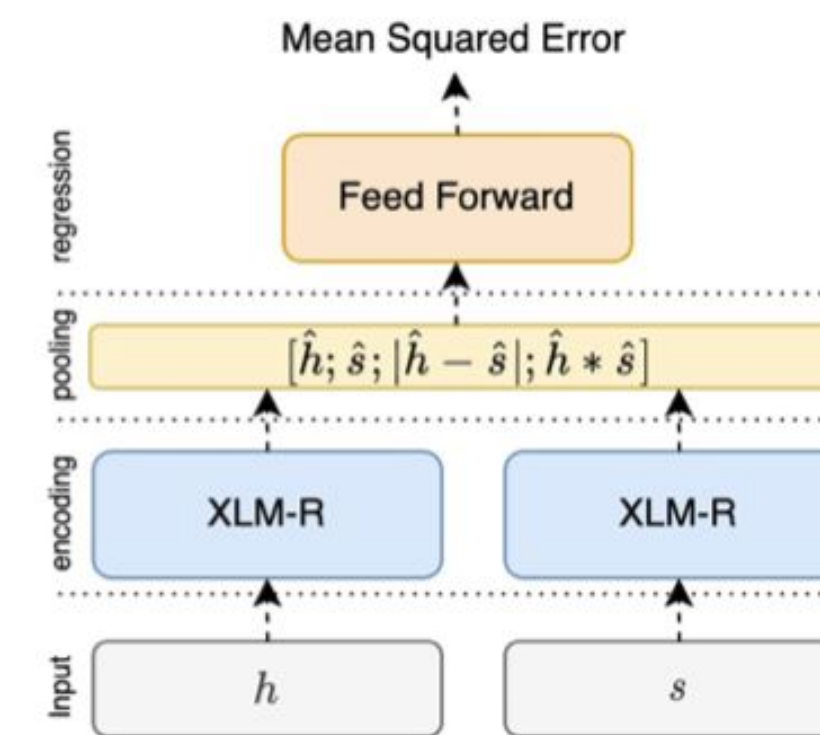
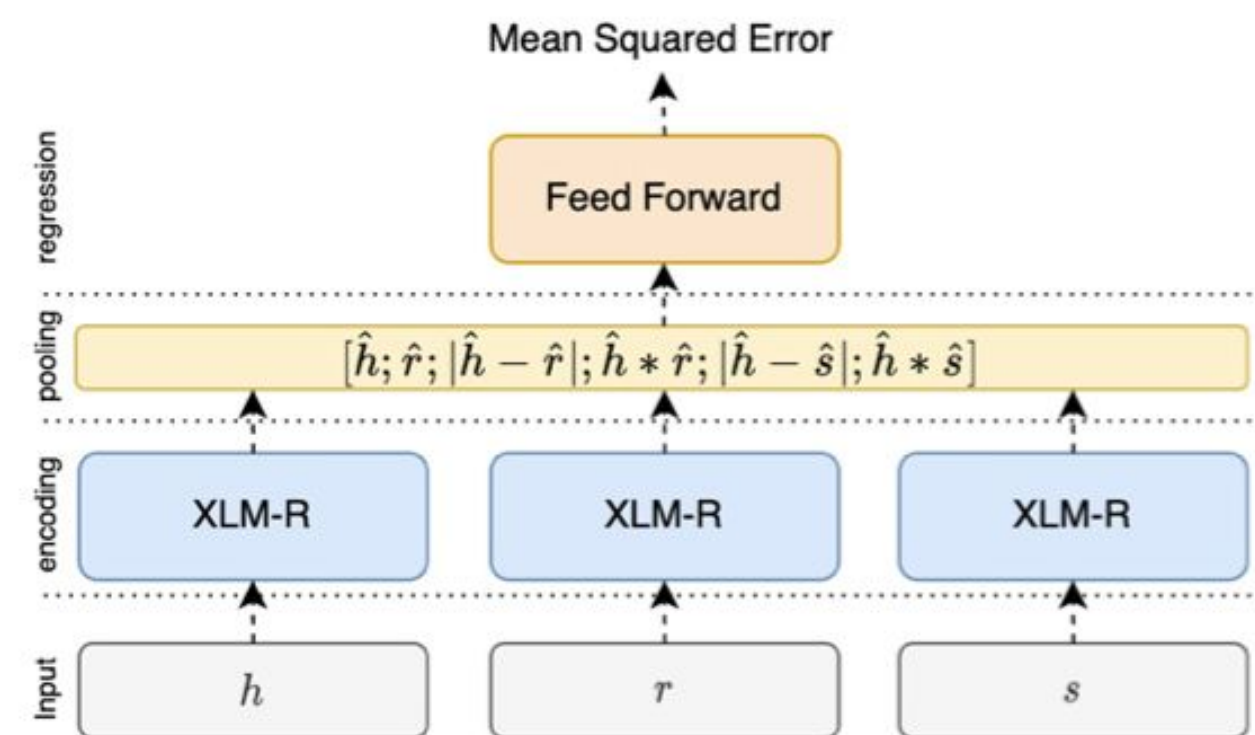
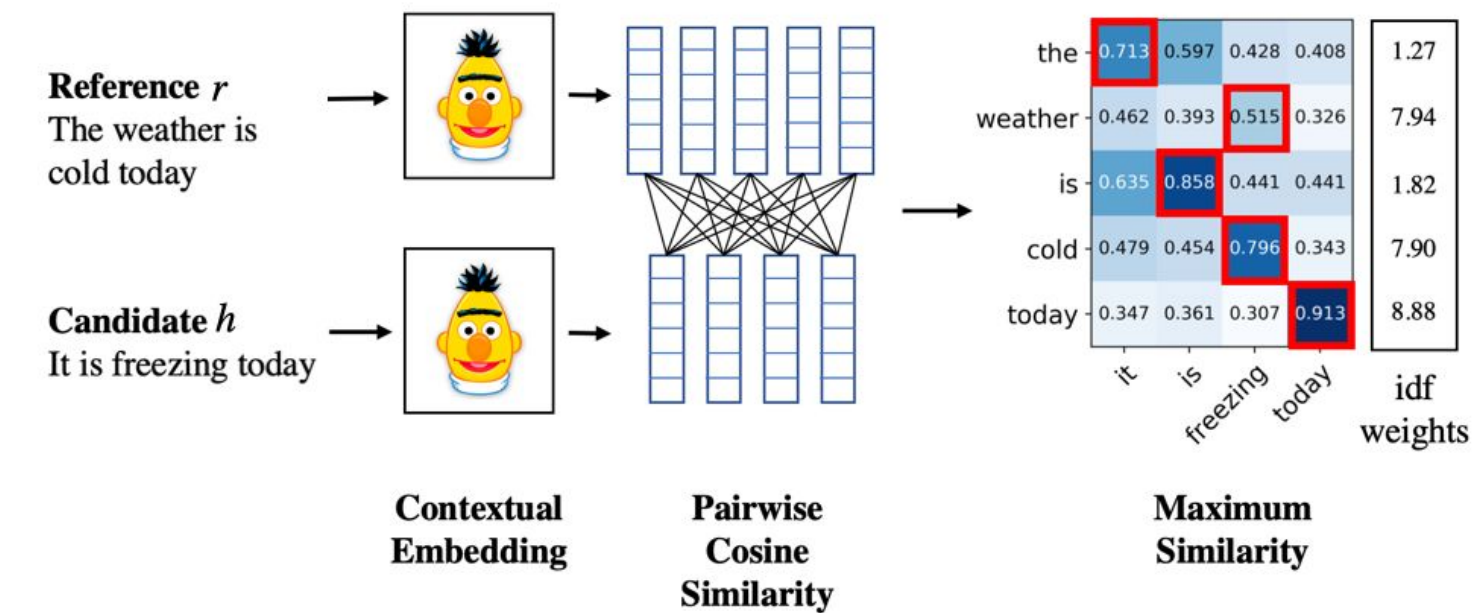
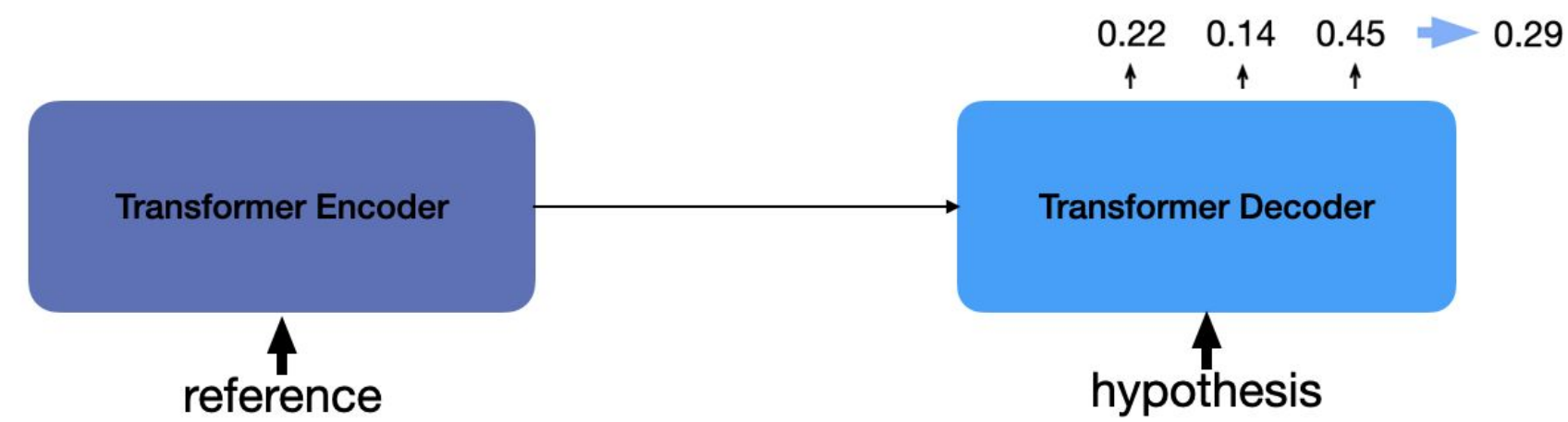
How accurately does the **entire** candidate document in German (deutsch) (right column) convey the original semantics of the source document in English (left column)?

Reset Submit

Appraise interface from [9].

Problem Formulation

Overlap-based and learned metrics still operate on the **sentence-level**



How can we incorporate context into learned metrics?

Related Work

Document-level context has also been proven useful for MT systems

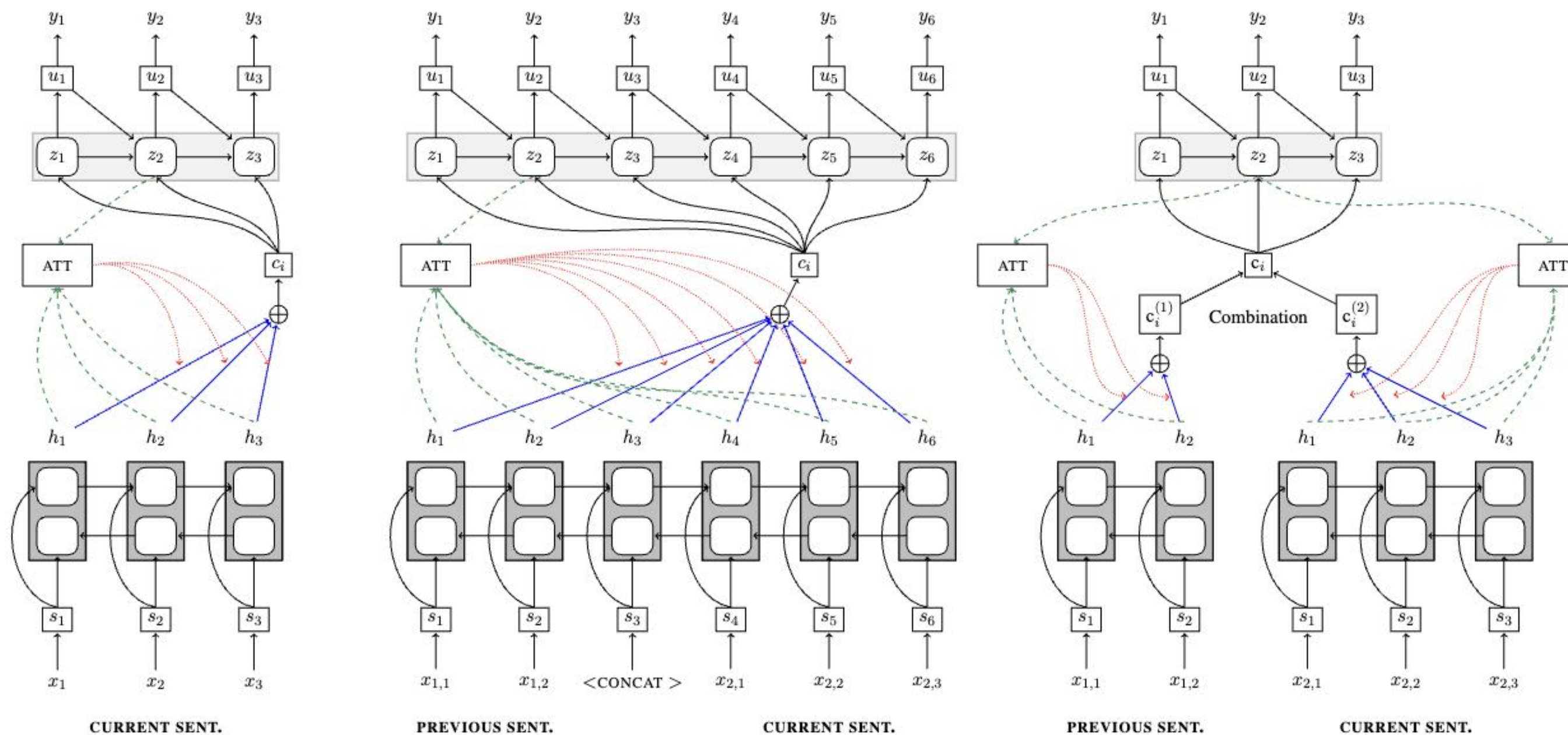


Figure from [12].

- Different ways to encode context: concatenation, encoders, gating
- Unclear if translation quality improves: human evaluation or targeted datasets[14]

Related Work

Context usage is mostly unexplored in automatic MT metrics

		ENTITY \mathcal{E}	TENSE \mathcal{V}	PRONOUN \mathcal{P}	DM \mathcal{M}
SRC	a) 小乔(Qiao) 看着(look) 相片回忆(recall) 起了二十年前。 b) 那个满脸胡须的男人(man) 正是(be)她(she) 的新婚丈夫。 c) 那却是(be) 他们之间初次见面(meet)。 d) 小乔(Qiao)一见到他(he) 心里就咯噔(jolt) 了一下, 噌的站(stand) 起来。	[[Qiao]]	[VBD, VBZ]	[masculine, feminine, epicene, neuter]	[contiguency, temporal, expansion, comparison]
REF	a) Qiao looked at the photo and recalled twenty years ago. b) This bearded man was her newlywed husband, c) [yet] this was the first time they were meeting with each other. d) [So] Qiao's heart jolted as soon as [she] saw him, and [she] quickly stood up.	[1] [0] [0] [1]	[2, 0] [1, 0] [2, 0] [2, 0]	[0, 0, 0, 0] [0, 1, 0, 0] [0, 0, 1, 0] [1, 2, 0, 0]	- [0, 0, 0, 0] [0, 0, 0, 1] [1, 0, 0, 0]
MTA	a) Qiao looked at the photo and recalled twenty years ago. b) This bearded man is her newlywed husband. c) This is the first time they meet with each other. d) Joe's heart is squeaky as soon as [he] saw him, and [he] quickly stands up.	[1] [0] [0] [0]	[2, 0] [0, 1] [0, 2] [0, 2]	[0, 0, 0, 0] [0, 1, 0, 0] [0, 0, 1, 0] [3, 1, 0, 0]	- [0, 0, 0, 0] [0, 0, 0, 0] [0, 0, 0, 0]
MTB	a) Qiao looked at the photo and recalled the past twenty years ago. b) This man with the beard was her newly-wed husband. c) [However], that was the first time they met. d) [So] as soon as Qiao saw him, [her] heart became squeaky, and [she] swiftly stood up.	[1] [0] [0] [1]	[2, 0] [1, 0] [2, 0] [2, 0]	[0, 0, 0, 0] [0, 1, 1, 0] [0, 0, 1, 0] [1, 2, 0, 0]	- [0, 0, 0, 0] [0, 0, 0, 1] [1, 0, 0, 0]

BlonDe: an overlap-based document-level MT metric for English that focuses on discourse phenomena^[15]

Document-level MT Metrics

Simple and **effective** approach -> add **context** during inference:

Document-level MT Metrics

Simple and **effective** approach -> add **context** during inference:

Instead of just using source hypothesis and reference concatenate source, hypothesis and reference context

Document-level MT Metrics

Simple and **effective** approach -> add **context** during inference:

Instead of just using source hypothesis and reference concatenate source, hypothesis and reference context

- No retraining
- No document-level human annotations

Document-level MT Metrics

Simple and **effective** approach -> add **context** during inference:

Instead of just using source hypothesis and reference concatenate source, hypothesis and reference context

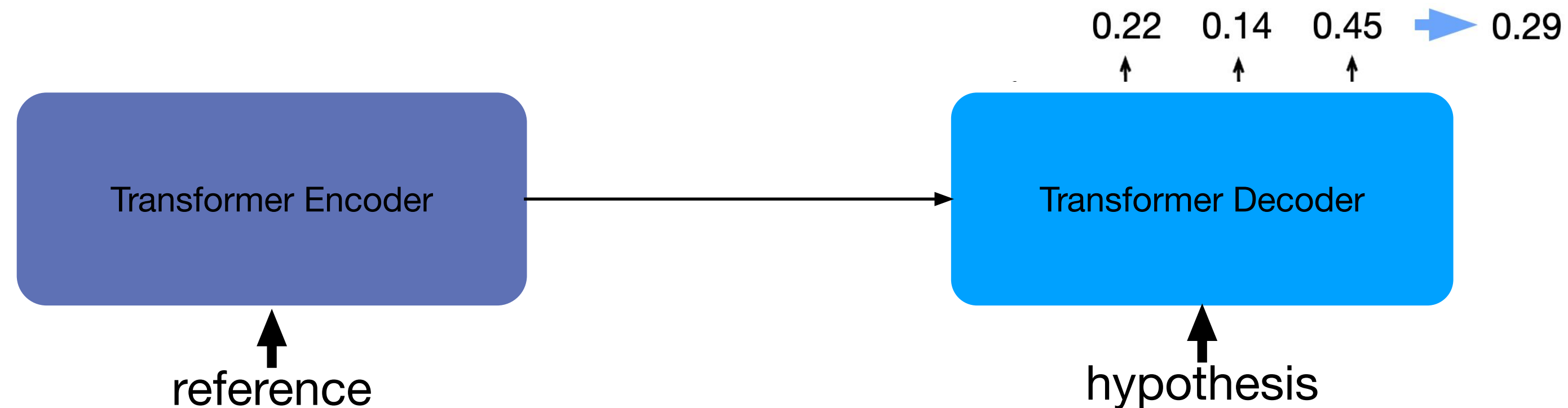
- No retraining
- No document-level human annotations

*Score **one sentence** at a time using document-level context*

Document-level MT Metrics

Simple and **effective** approach -> add **context** during inference:

Prism

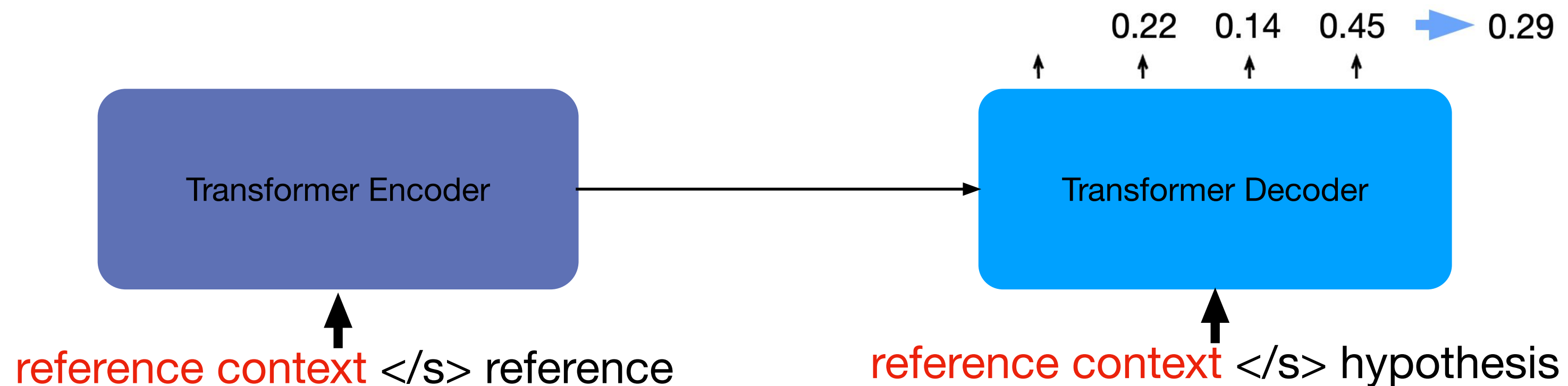


- Test whether the hypothesis is a paraphrase of the reference and vice versa
- A multilingual MT model that was trained at the sentence level as the paraphrase model (m39v1)

Document-level MT Metrics

Simple and **effective** approach -> add **context** during inference:

Document-level Prism

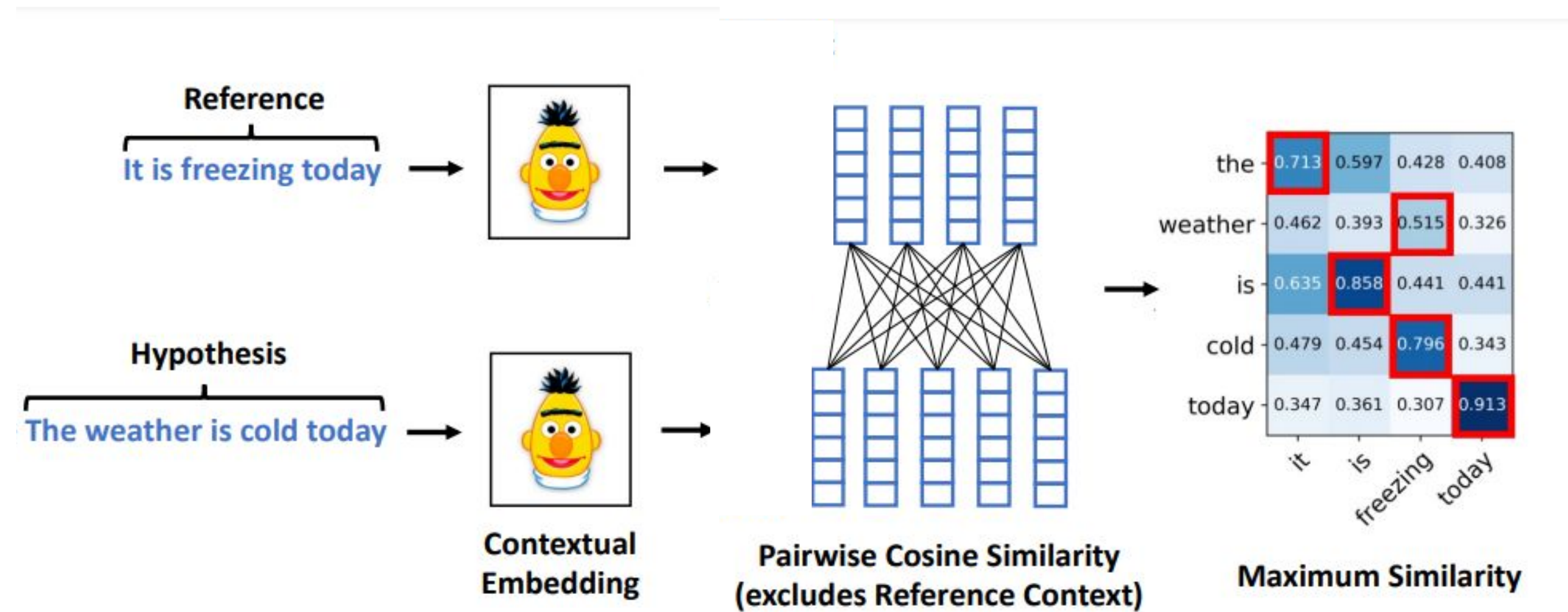


- We use mBART-50_[16] a multilingual LM that was trained at the document level as the paraphrase model
- We concatenate the **reference context** to both the encoder and decoder
- We only compute token-level probabilities for the **sentence** we want to score

Document-level MT Metrics

Simple and effective approach -> add **context** during inference:

BERTScore

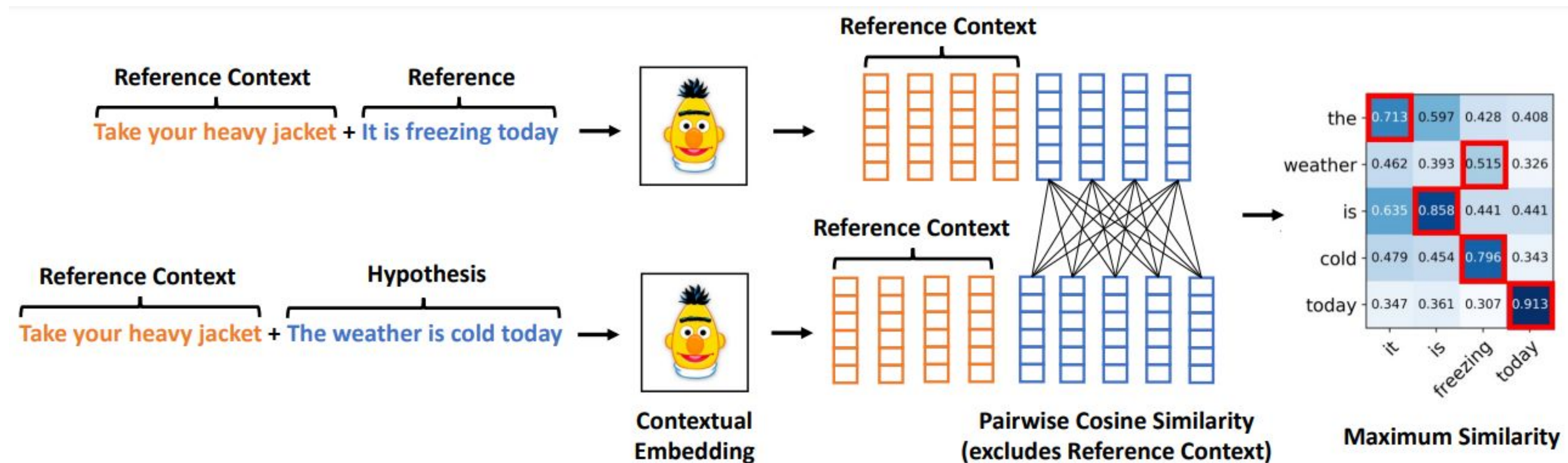


- Contextual embeddings from BERT
- Soft-alignment between words
- Greedy matching from matrix to calculate precision, recall and F1

Document-level MT Metrics

Simple and effective approach -> add **context** during inference:

Document-level BERTScore

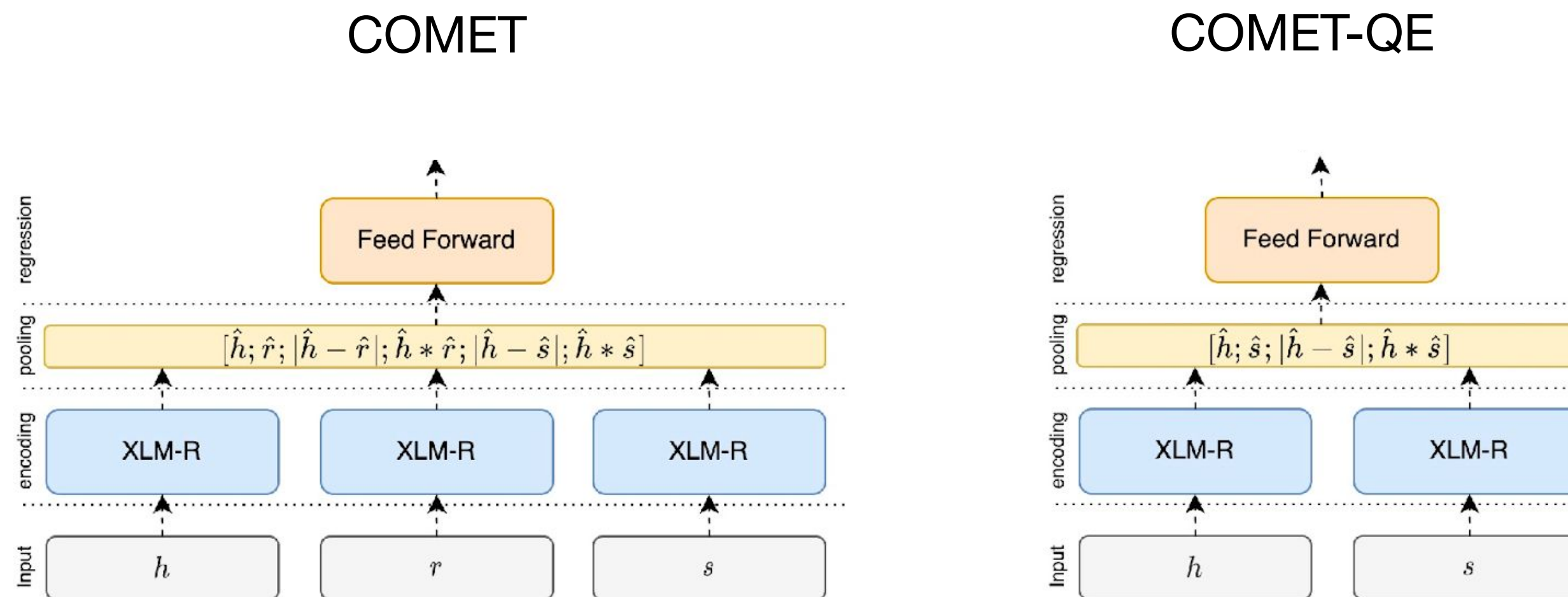


- We concatenate the **reference context** when encoding the reference or the hypothesis with the LM
- We only align the tokens of the **current sentence** by setting all the other similarity scores to zero
- BERT is pretrained on chunks of text (512 tokens)

Document-level MT Metrics

Simple and **effective** approach -> add **context** during inference:

COMET

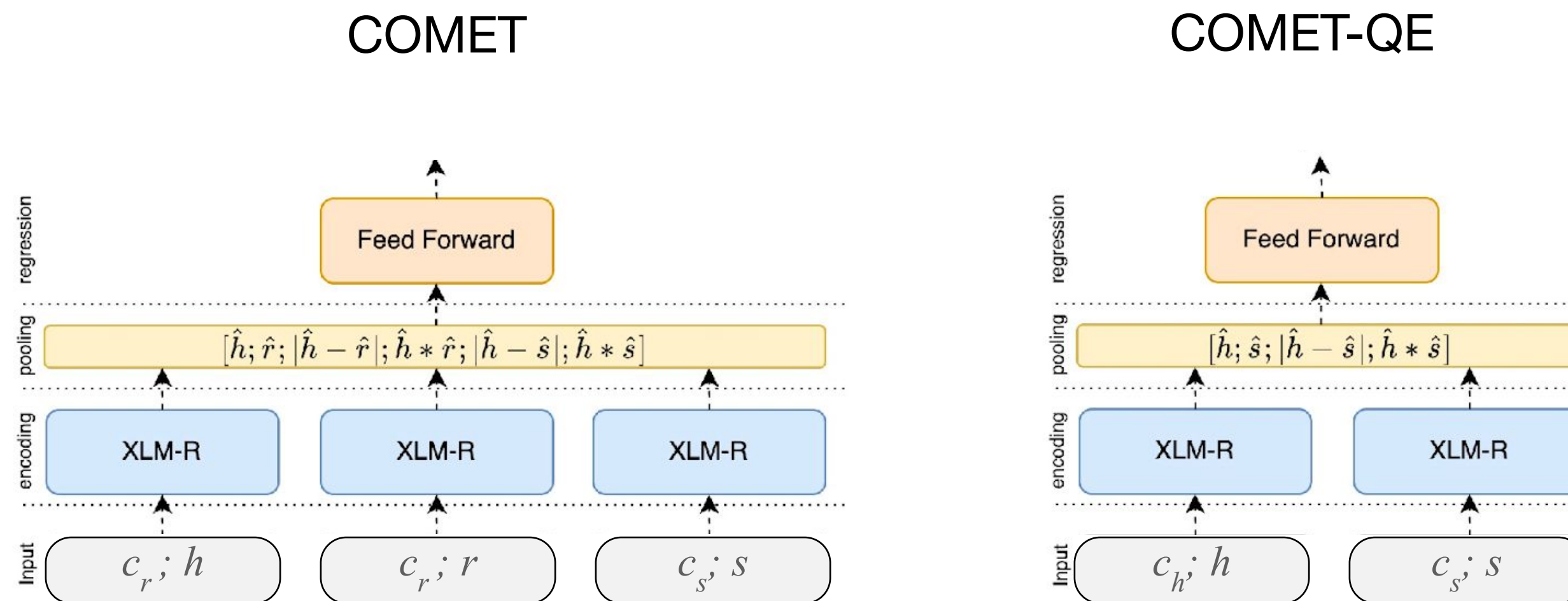


- Contextual embeddings from XLM-R_[16] for source, candidate and reference
- Average pooling of output token embeddings
- Model is trained to predict human scores

Document-level MT Metrics

Simple and **effective** approach -> add **context** during inference:

Document-level COMET



- We concatenate **source** and **reference context** with the **source** and **reference** sentences in the encoder
- We average the embeddings of the current sentence only
- XLM-R is pretrained on chunks of text

Experiments

We evaluate our approach on the MQM annotations of WMT21 Metrics Task[18]:

- MQM guidelines strongly advise annotators to take context into account[11]
- Two different domains, News (articles, long sentences) and TED talks (transcribed speech, shorter sentences, contextual phenomena)

All our models can handle more than one sentence as input.

We use the *two previous sentences* as context.

We substitute the hypothesis context, c_h , with the reference context, c_r , when available to avoid propagation of errors.

Results

Model	Input	TED talks			News		
		En→De	En→Ru	Zh→En	En→De	En→Ru	Zh→En
BlonDe	$\langle c_h, h, c_r, r \rangle$	-	-	-0.232	-	-	0.212
Prism (m39v1)	$\langle h, r \rangle$	0.656	0.867	0.272	0.841	0.799	0.558
Prism (mBART-50)	$\langle h, r \rangle$	0.486	0.845	0.240	0.661	0.710	0.363
Doc-Prism (mBART-50)	$\langle c_r; h, c_r; r \rangle$	0.692	0.852	0.372	0.825*	0.777	0.374
BERTScore	$\langle h, r \rangle$	0.506	0.831	0.293	0.930	0.629	0.575*
Doc-BERTScore	$\langle c_r; h, c_r; r \rangle$	0.613*	0.836	0.344*	0.948*	0.622	0.535
COMET	$\langle s, h, r \rangle$	0.818	0.841	0.266	0.772	0.659	0.628
Doc-COMET	$\langle c_s; s, c_r; h, c_r; r \rangle$	0.816	0.849	0.297	0.802*	0.676	0.513
COMET-QE	$\langle s, h \rangle$	0.694	0.818	-0.209	0.711	0.688	0.529
Doc-COMET-QE	$\langle c_s; s, c_h; h \rangle$	0.724	0.830	-0.255	0.733	0.733*	0.462

System-level Pearson correlation with WMT21 MQM annotations for the news domain and TED talks.
Results for baselines and trained metrics with (Doc-*) and without context.

Results

Model	Input	TED talks			News		
		En→De	En→Ru	Zh→En	En→De	En→Ru	Zh→En
BlonDe	$\langle c_h, h, c_r, r \rangle$	-	-	-0.232	-	-	0.212
Prism (m39v1)	$\langle h, r \rangle$	0.656	0.867	0.272	0.841	0.799	0.558
Prism (mBART-50)	$\langle h, r \rangle$	0.486	0.845	0.240	0.661	0.710	0.363
Doc-Prism (mBART-50)	$\langle c_r; h, c_r; r \rangle$	0.692	0.852	0.372	0.825*	0.777	0.374
BERTScore	$\langle h, r \rangle$	0.506	0.831	0.293	0.930	0.629	0.575*
Doc-BERTScore	$\langle c_r; h, c_r; r \rangle$	0.613*	0.836	0.344*	0.948*	0.622	0.535
COMET	$\langle s, h, r \rangle$	0.818	0.841	0.266	0.772	0.659	0.628
Doc-COMET	$\langle c_s; s, c_r; h, c_r; r \rangle$	0.816	0.849	0.297	0.802*	0.676	0.513
COMET-QE	$\langle s, h \rangle$	0.694	0.818	-0.209	0.711	0.688	0.529
Doc-COMET-QE	$\langle c_s; s, c_h; h \rangle$	0.724	0.830	-0.255	0.733	0.733*	0.462

System-level Pearson correlation with WMT21 MQM annotations for the news domain and TED talks.
Results for baselines and trained metrics with (Doc-*) and without context.

- Significantly outperform document-level metric baseline
- Improved correlation for TED talks across metrics and language pairs
- Improvements on news for 2/3 pairs.
Zh->En reference is of lower quality (MQM score 4.2, best MT 4.47)[19]

Results

Q: *Do the gains of our approach come from contextual phenomena?*

We use contrastive sets (ContraPro^[20], DiscEvalMT^[12]) used to evaluate document-level MT models.

source:	<i>It could get tangled in your hair.</i>
reference:	<i>Sie könnte sich in deinem Haar verfangen.</i>
contrastive:	<i>Er könnte sich in deinem Haar verfangen.</i>
contrastive:	<i>Es könnte sich in deinem Haar verfangen.</i>

We consider the original and contrastive references as outputs of different MT systems.

We only test reference-free metrics (COMET-QE), otherwise the task is trivial.

Results

Q: *Do the gains of our approach come from contextual phenomena?*

We use contrastive sets (ContraPro^[20], DiscEvalMT^[12]) used to evaluate document-level MT models.

source:	<i>It could get tangled in your hair.</i>
reference:	<i>Sie könnte sich in deinem Haar verfangen.</i>
contrastive:	<i>Er könnte sich in deinem Haar verfangen.</i>
contrastive:	<i>Es könnte sich in deinem Haar verfangen.</i>

Model	En → De			En → Fr				
	Intra	Inter	Total	Intra	Inter	Total	Anaphora	Disambiguation
Doc-MT (Lopes et al., 2020)	-	-	70.8	-	-	83.2	82.5	55.0
COMET-QE	78.2	40.9	48.4	76.3	76.6	76.5	50.0	50.0
Doc-COMET-QE (this work)	80.5	72.6	74.2	88.7	88.0	88.3	83.5	68.0

Accuracy (%) for targeted evaluation of contextual phenomena.

- Consistent and significant improvements using context
- Our approach outperforms document-level MT systems
- Gains even when the antecedent is in the current sentence (Intra)

Results

Q: *Do the gains of our approach come from contextual phenomena?*

A: **YES**

We use contrastive sets (ContraPro^[20], DiscEvalMT^[12]) used to evaluate document-level MT models.

source:	<i>It could get tangled in your hair.</i>
reference:	<i>Sie könnte sich in deinem Haar verfangen.</i>
contrastive:	<i>Er könnte sich in deinem Haar verfangen.</i>
contrastive:	<i>Es könnte sich in deinem Haar verfangen.</i>

Model	En → De			En → Fr				
	Intra	Inter	Total	Intra	Inter	Total	Anaphora	Disambiguation
Doc-MT (Lopes et al., 2020)	-	-	70.8	-	-	83.2	82.5	55.0
COMET-QE	78.2	40.9	48.4	76.3	76.6	76.5	50.0	50.0
Doc-COMET-QE (this work)	80.5	72.6	74.2	88.7	88.0	88.3	83.5	68.0

Accuracy (%) for targeted evaluation of contextual phenomena.

- Consistent and significant improvements using context
- Our approach outperforms document-level MT systems
- Gains even when the antecedent is in the current sentence (Intra)

Analysis

Q: *Does context quality play an important role?*

We substitute the hypothesis context, c_h , with the reference context, c_r , in all metrics but COMET-QE.

	Context	Doc-Prism	Doc-BERTScore	Doc-COMET
hypothesis	$\langle c_s; s, c_r; r, c_h; h \rangle$	0.595	0.624	0.630
reference	$\langle c_s; s, c_r; r, c_r; h \rangle$	0.649	0.650	0.659

Average correlation for all domains and language pairs using hypothesis vs reference context.

- Hypothesis context leads indeed to worse correlation
- Conditioning on low-quality context has diminishing results (e.g. Zh->En)

Analysis

Q: *Does context quality play an important role?*

A: **YES**

We substitute the hypothesis context, c_h , with the reference context, c_r , in all metrics but COMET-QE.

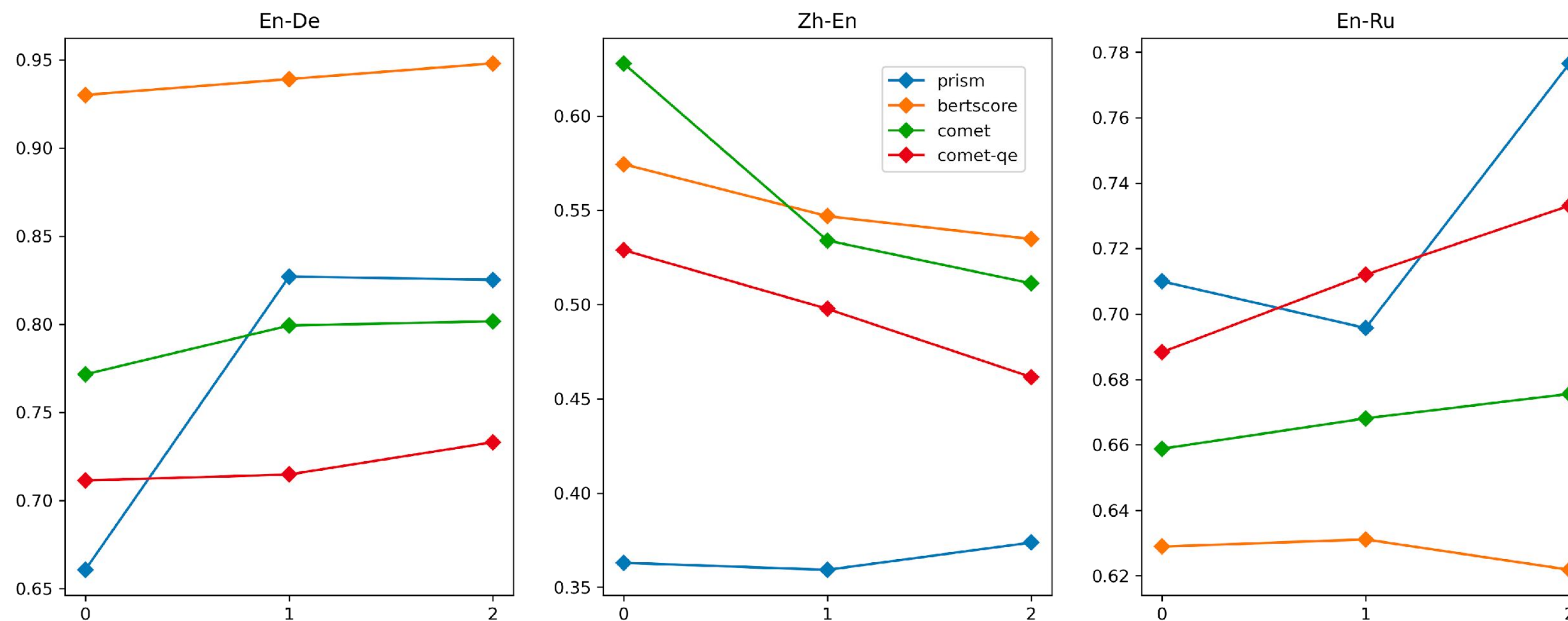
	Context	Doc-Prism	Doc-BERTScore	Doc-COMET
hypothesis	$\langle c_s; s, c_r; r, c_h; h \rangle$	0.595	0.624	0.630
reference	$\langle c_s; s, c_r; r, c_r; h \rangle$	0.649	0.650	0.659

Average correlation for all domains and language pairs using hypothesis vs reference context.

- Hypothesis context leads indeed to worse correlation
- Conditioning on low-quality context has diminishing results (e.g. Zh->En)

Analysis

Q: *How much context should be used in document-level MT metrics?*



Correlation vs. amount of context for news articles.

- Adding more context helps for 2/3 pairs
- For Zh->En using less context helps, due to low quality of the reference

Conclusion

- **Simple and effective** approach towards document-level MT metrics
- **No retraining or additional data** needed
- **Consistent improvements** across all metrics (TED talks)
- Gains come from **better context utilization**

Conclusion

- **Simple and effective** approach towards document-level MT metrics
- **No retraining or additional data** needed
- Consistent **improvements** across all metrics (TED talks)
- Gains come from **better context utilization**

Limitations

- Not fully **document-level** : consistency, fluency
- **Context** might be **redundant** in some cases

Future Work

- Explore other ways of integrating context (e.g. gating)
- Retrain/Adapt metrics on document-level annotations

Thank you! Questions?

COMET PR (work in progress)

Scoring MT outputs:

Command Line usage:

To score using the original, sentence-level COMET/COMET-QE models:

```
comet-score -s src.de -t hyp1.en -r ref.en --model wmt21-comet-mqm  
comet-score -s src.de -t hyp1.en --model wmt21-comet-qe-mqm
```

To score using the document-level COMET/COMET-QE simply add the `--doc` flag:

```
comet-score -s src.de -t hyp1.en -r ref.en --doc --model wmt21-comet-mqm  
comet-score -s src.de -t hyp1.en --doc --model wmt21-comet-qe-mqm
```

References

- [1] *Bleu: a Method for Automatic Evaluation of Machine Translation*, Papineni et al., 2002, ACL
- [2] *chrF: character n-gram F-score for automatic MT evaluation*, Popovic et al., 2015, Workshop on Statistical Machine Translation
- [3] *A study of translation edit rate with targeted human annotation*, Snover et al., 2006, AMTA
- [4] *BERTScore: Evaluating text generation with BERT*, Zhang et al., 2020, ICLR
- [5] *COMET: A neural framework for MT evaluation*, Rei et al., 2020, EMNLP
- [6] *Automatic machine translation evaluation in many languages via zero-shot paraphrasing*, Thompson et al., 2020, EMNLP
- [7] *Results of the WMT20 metrics shared task*, Mathur et al., 2020, WMT
- [8] *On Context Span Needed for Machine Translation Evaluation*, Castilho et al., 2020, LREC
- [9] *Findings of the 2020 Conference on Machine Translation (WMT20)*, Barault et al., 2020, WMT
- [10] *Has machine translation achieved human parity? a case for document-level evaluation*, Läubli et al., 2018, EMNLP
- [11] *Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation*, Freitag et al., 2021, TACL
- [12] *Evaluating Discourse Phenomena in Neural Machine Translation*, Bawden et al., 2018, NAACL

References

- [13] *Document-level Neural MT: A Systematic Comparison*, Lopes et al., 2020, EAMT
- [14] *chrF: character n-gram F-score for automatic MT evaluation*, Popovic et al., 2015, Workshop on Statistical Machine Translation
- [15] *BlonDe: An automatic evaluation metric for document-level machine translation*, Jiang et al., 2022, NAACL
- [16] *Multilingual Translation with Extensible Multilingual Pretraining and Finetuning*, Tang et al., 2020, Arxiv
- [17] *Unsupervised Cross-lingual Representation Learning at Scale*, Conneau et al., 2020, ACL
- [18] *Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain*, Freitag et al., 2021, WMT
- [19] *Findings of the 2021 Conference on Machine Translation (WMT21)*, Akhbardeh et al., 2021, WMT
- [20] *A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation*, Müller et al., 2018, WMT