DATScore: Evaluating Translation with Data Augmented Translation

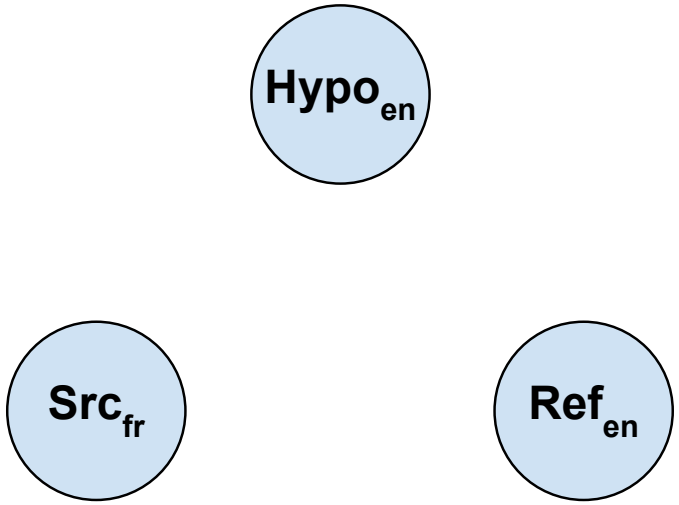
Moussa Kamal Eddine[1], Guokan Shang[2], Michalis Vazirgiannis[1,3]
[1]École Polytechnique; [2]Linagora; [3]AUEB

EACL 2023, Dubrovnik, Croatia, 1st – 6th May 2023
Contact: moussa.kamal-eddine@polytechnique.edu
Paper: https://arxiv.org/abs/2210.06576

LaTeXof the slides: https://www.overleaf.com/read/ccwyvpwvhzmc

Introduction (1/3)

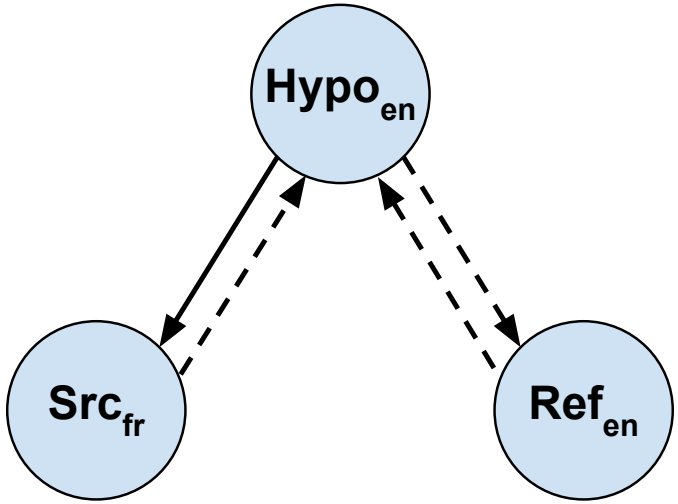
source: original text, **reference**: human translation
hypothesis: system-generated translation



reference-free: Hypo vs. Src
reference-based: Hypo vs. Ref (e.g., BLEU, BERTScore, MoverScore ⇒ match tokens or their embeddings.)

Introduction (2/3)

source: original text, **reference**: human translation
hypothesis: system-generated translation



BARTScore proposed a novel conceptual view ⇒ It treats *the evaluation of generated text as a text generation problem*.
BARTScore directly uses BART model's conditional probability of generating a provided target text Y given a provided input text X, as the evaluation score of the generation direction X → Y.

Introduction (3/3)

One Direction's Score

The score for the generation direction from a source sequence $X = \{x_t\}_{t=1}^{n}$ to a target sequence $Y = \{y_t\}_{t=1}^{m}$ is calculated as the factorized, weighted log probability over all generation steps:

$$\text{Score}_{X \to Y} = \sum_{t=1}^{m} w_t \log P(y_t | X, \{y_{t'}\}_{t'=1}^{t-1}; \theta) \quad (1)$$

- w_t denotes the term importance score to put different emphasis on different target tokens y_t.
  BARTScore simply employs a uniform weighting scheme (all equal to 1).
- θ denotes the BART model.

Our Approach (1/4)

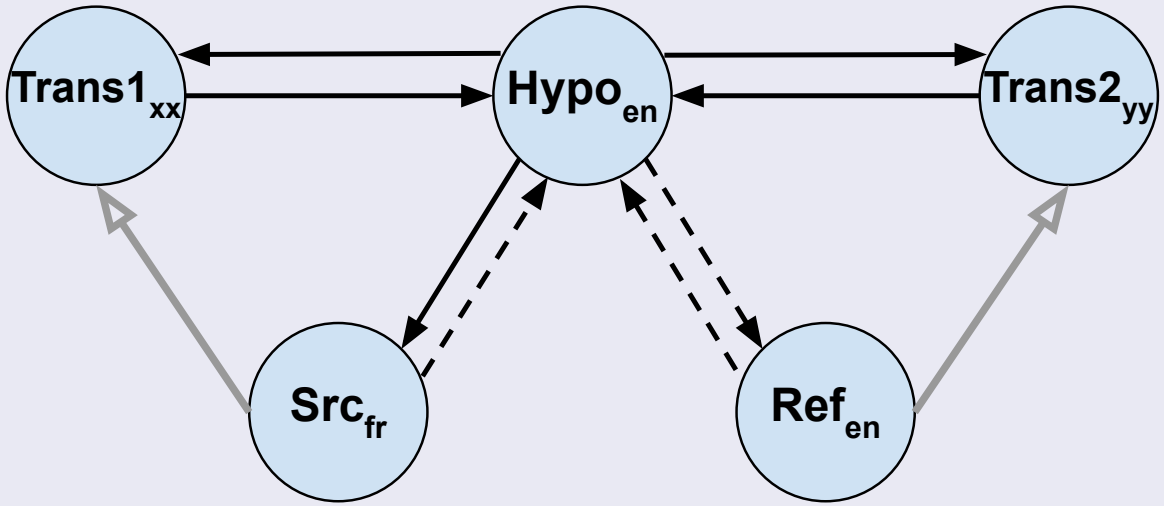


Figure: An illustration of the directions used in the final aggregated DATScore.

- Trans1$_{xx}$ and Trans2$_{yy}$ represent data *augmented translations* in any languages xx and yy
- To estimate this probability we use pretrained multilingual MT model (M2M-100 in our case)

Our Approach (2/4)

DATScore

DATScore is calculated as the weighted average of the scores associated with all the directions:

$$\text{DATScore} = \sum_{X,Y} w_{X \to Y} \text{Score}_{X \to Y}; X \neq Y \quad (2)$$

where $w_{X \to Y}$ denotes the weight of the direction $X \to Y$.

Our Approach (3/4)

One-vs-rest score averaging method.

- One direction score might strongly disagree with the others, likely being an outlier.
- Each direction is weighted with the sum of the Pearson correlations of its scores with the scores of all the other directions.

$$w_{X \to Y} = \sum_{X', Y'} Corr(\text{Score}_{X \to Y}, \text{Score}_{X' \to Y'}) \quad \text{s.t.} \quad (X, Y) \neq (X', Y') \quad (3)$$

Our Approach (4/4)

Entropy-based term weighting scheme.

The assumption is that when the model is very confident in generating the target token (low entropy), then this token is non-informative (e.g., stopword).

$$w_t = -\sum_{i=1}^{v} \mathrm{P_t}(z_i) \log \mathrm{P_t}(z_i) \tag{4}$$

where v denotes the size of the output generation vocabulary. $\mathrm{P_t}(z_i)$ represents the probability of the i-th token in the vocabulary at time step t.

Results (1/2)

Metric		Model	$\|r\|$:cs → en / τ:en → cs	$\|r\|$:de → en / τ:en → de	$\|r\|$:fi → en / τ:en → fi	$\|r\|$:lv → en / τ:en → lv	$\|r\|$:ru → en / -	$\|r\|$:tr → en / τ:en → tr	$\|r\|$:zh → en / -	Avg.
BLEU	1a)	N/A	34.4/22.0	36.6/23.6	44.4/42.1	32.1/21.5	41.3/-	44.1/33.6	44.0/-	37.8/27.3
BERTScore	1b)	RL/mBERT	71.0/43.8	**74.5**/40.4	83.3/58.8	75.6/46.6	74.6/-	75.1/57.1	77.5/-	75.9/49.3
MoverScore	1c)	BB/mBERT	66.6/38.3	70.6/35.9	82.2/54.2	71.7/37.8	73.7/-	76.1/49.8	74.3/-	73.6/43.2
BARTScore	1d)	BL+para/mBART	68.4/39.0	70.8/33.4	79.4/50.4	74.9/50.4	71.8/-	73.9/53.8	76.0/-	73.6/45.4
	1e)	M2M-100_418M	65.9/45.0	66.1/44.5	79.9/59.2	71.7/40.3	69.0/-	71.8/70.9	71.6/-	70.9/52.0
	1f)	M2M-100_1.2B	67.4/49.6	69.3/49.2	80.7/63.5	73.7/46.9	70.4/-	71.6/**72.5**	73.0/-	72.3/56.3
DATScore	1g)	M2M-100_418M	68.6/51.1	68.5/48.1	82.0/63.7	74.7/48.3	73.0/-	77.6/70.9	76.5/-	74.4/56.4
	1h)	M2M-100_1.2B	**71.3**/**53.9**	72.9/**52.2**	**83.5**/**66.3**	**76.8**/**52.0**	**75.9**/-	**78.1**/70.9	**77.7**/-	**76.6**/**59.1**

Table: Absolute Pearson correlation ($|r|$) for `to-English` and Kendall correlations (τ) for `from-English` with segment-level human scores on WMT17. BB stands of Bert-Base, RL for RoBERTa-Large and BL for BART-Large.

$\Rightarrow$ our metric provides a performance boost of 0.7 for to-English case and of 9.8 for from-English case on WMT17 dataset (v.s. 1b)

Results (2/2)

Metric		Model	τ:cs → en / τ:en → cs	τ:de → en / τ:en → de	τ:et → en / τ:en → et	τ:fi → en / τ:en → fi	τ:ru → en / τ:en → ru	τ:tr → en / τ:en → tr	τ:zh → en / τ:en → zh	Avg.
BLEU	2a)	N/A	23.3/38.9	41.5/62.0	38.5/41.4	15.4/35.5	22.8/33.0	14.5/26.1	17.8/31.1	24.8/38.3
BERTScore	2b)	RL/mBERT	40.4/55.9	**55.0**/72.7	39.7/58.4	29.6/53.9	35.3/42.4	29.2/38.9	**26.4**/36.1	36.5/51.2
MoverScore	2c)	BB/mBERT	36.8/44.6	53.9/68.4	39.4/52.7	28.7/50.9	27.9/40.1	**33.6**/32.5	25.6/35.2	35.1/46.3
BARTScore	2d)	BL+para/mBART	39.6/50.2	54.7/65.0	39.4/53.3	28.9/57.2	34.6/37.0	27.4/37.7	24.9/32.4	35.6/47.5
	2e)	M2M-100_418M	36.3/55.4	53.5/72.2	37.6/58.4	26.3/60.2	33.4/44.4	26.8/45.1	23.4/31.3	33.9/52.4
	2f)	M2M-100_1.2B	38.4/**63.5**	54.6/**76.2**	39.2/63.2	27.9/64.5	35.7/45.6	28.5/50.2	24.3/34.7	35.5/56.8
DATScore	2g)	M2M-100_418M	38.6/53.5	53.5/71.3	39.3/64.0	28.4/62.2	34.9/44.4	28.5/47.9	25.3/34.0	35.5/53.9
	2h)	M2M-100_1.2B	**40.7**/61.9	54.9/**76.2**	**40.5**/**68.2**	**30.4**/**67.9**	**36.4**/**46.2**	31.0/**52.7**	26.3/**36.6**	**37.2**/**58.5**

Table: Kendall correlations (τ) for `to-English` and `from-English` with segment-level human scores on WMT18. BB stands of Bert-Base, RL for RoBERTa-Large and BL for BART-Large.

$\Rightarrow$ our metric achieves a gain of of 0.7 for to-English case and of 7.3 for from-English case on WMT18 dataset (v.s. 2b)

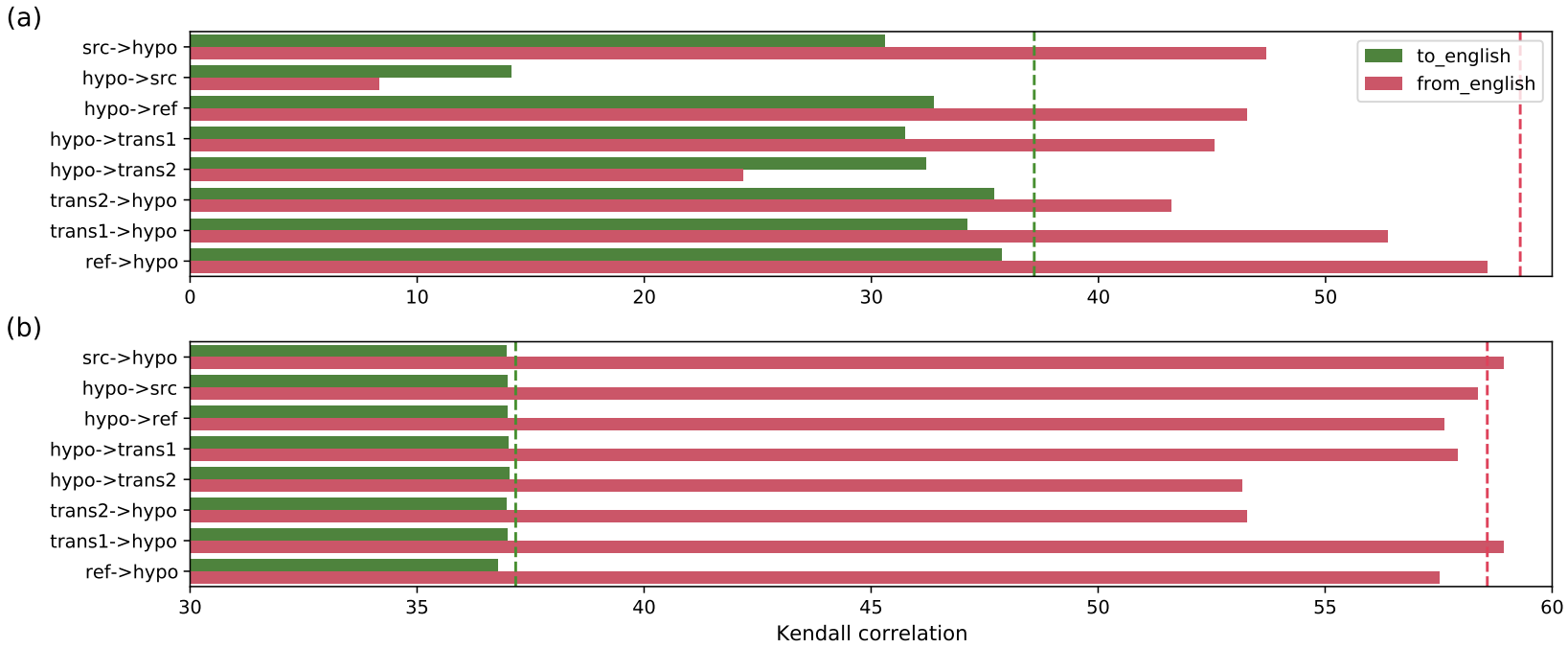
Results (3/3)

Metric	Model	WebNLG			REALSumm	SummEval				Flickr8K	PASCAL-50S
		SEMA	GRAM	FLU	COV	COH	CONS	FLU	REL	RELE	RR
BLEU	N/A	45.5	36.0	34.9	37.9	11.8	6.3	7.7	18.6	13.8	8.1
BERTScore	RoBERTa-Large	56.1	60.8	54.8	41.2	**33.9**	10.5	**15.0**	**35.9**	46.1	**33.8**
MoverScore	BERT-Base	-9.9	-27.8	-20.6	44.1	14.4	**14.7**	13.8	29.1	**52.5**	33.2
BARTScore	BART-Large+para	**71.9**	61.3	57.4	31.7	20.8	-3.5	6.7	22.2	44.8	33.1
	M2M-100_418M	64.9	62.8	56.0	30.1	14.8	-2.3	3.0	19.8	34.3	29.6
	M2M-100_1.2B	66.1	63.9	57.2	32.0	17.1	1.1	6.7	22.8	34.6	26.3
DATScore	M2M-100_418M	69.9	62.9	57.2	44.7	17.1	4.4	4.6	26.3	42.6	29.6
	M2M-100_1.2B	70.4	**63.7**	**57.9**	**45.5**	19.5	6.8	8.2	30.2	45.3	31.4

Table: Pearson correlation results on various NLG tasks: **Data-to-text** (WebNLG), **abstractive summarization** (REALSumm and SummEval), and **Image Captioning** (Flickr8K and PASCAL-50S).

$\Rightarrow$ although not the top-performing metric across all tasks, DATScore showed an overall stable and competitive performance.

Contributions of all direction scores.



Contribution of directions

- (a): The horizontal bars represent the Kendall correlations of **each individual generation direction**.
- (b): The horizontal bar represents the Kendall correlation of **a variant of DATScore with excluding the single generation direction** of the line.
- Both in (a) and (b), the dashed vertical lines represent the Kendall correlation of the vanilla and **complete DATScore**.

One-vs-rest and entropy-based weighting strategies.

Entropy-based weighting	One-vs-rest weighting	to_English	from_English
✓	✓	**37.2**	**58.5**
✓	✗	37.1	58.1
✗	✓	36.4	55.9
✗	✗	36.4	56.0

Table: The average Kendall correlation (to/from)-English when the entropy-based and one-vs-rest weighting are included or excluded. Experiments are conducted on WMT18.

Conclusion

Contributions

- We propose DATScore; an untrained and unsupervised translation evaluation metric that offers a large performance boost especially evaluating low-resource language generation.
- A novel one-vs-rest method to average the scores for different generation directions with different weights.
- A novel entropy-based scheme for weighting the target generated terms so that higher informative tokens receive more importance in accounting for the score.