IDA
2023

Diverse Paraphrasing with Insertion Models for Few-Shot Intent Detection

Raphaël Chevasson, Charlotte Laclau, and Christophe Gravier

https://github.com/RaphaelChevasson/DPIM

✉ raphael.chevasson@univ-st-etienne.fr

Overview

I – Motivation

AR and control, alternatives

II – Method

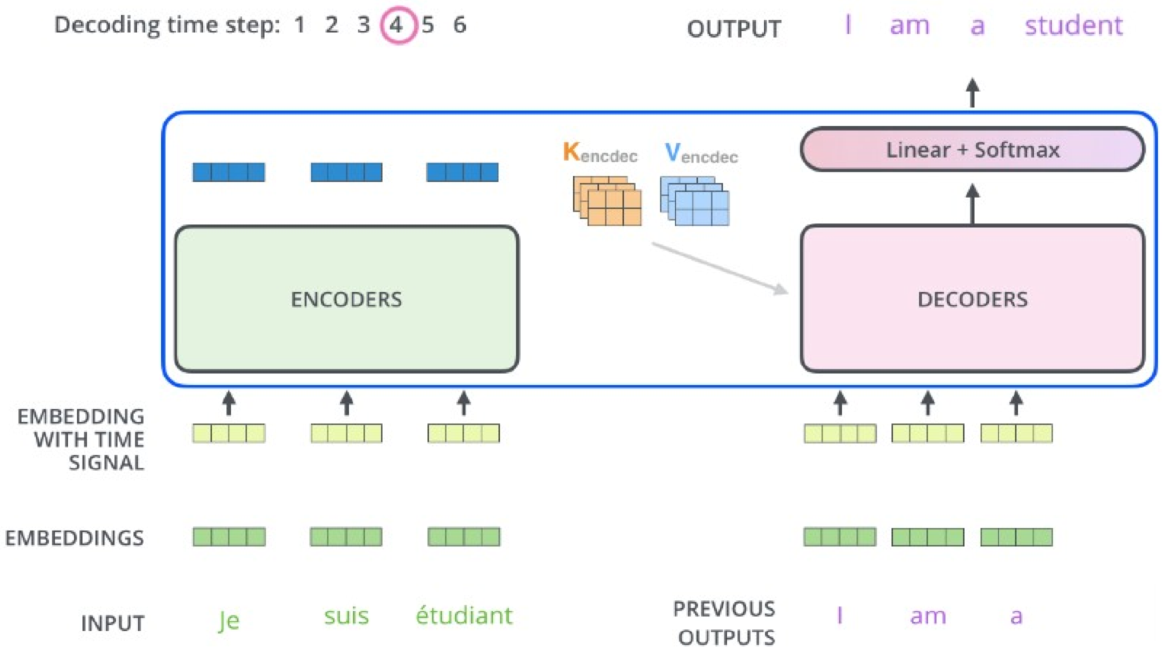
a) paraphrasing pipeline
b) metrics
c) datasets

III – Results & perspectives

a) 3-way tradeoff
b) qualitative and sample analysis

Scientific context

Current Text Generation is dominated by Autoregressive, decoder-only transformers:



Task

Artificial Intelligence
→ NLP (tokens)
→ Text Generation

Method

Deep Learning
→ Transformers
→ Autoregressive

[Alammar 2018] Jay Alammar: The Illustrated Transformer, June 27, 2018, https://jalammar.github.io/illustrated-transformer/

Motivation

Current NLG dominated by AR, decoder-only transformers

(+) can train every token in one pass with masked attention

(+) scale well across data and compute

Controlled generation can be hard

(-) cannot force a token to be included

(-) cannot force both left and right context

Motivation

Alternatives :

Encoder-decoder

→ can solve the bidirectionality (e.g. XLNet [Yang 2019])

Non autoregressive

→ one-pass, lower quality generation (e.g. the Non-Autoregressive Transformer [Gu 2018])

Semi-autoregressive

→ lots of very diverse works (cf the NAR tutorial [Gu 2022] from ACL 2022 for a very good overview)

[Gu 2018] Jiatao Gu, et. al.: Non-Autoregressive Neural Machine Translation, ICLR 2018

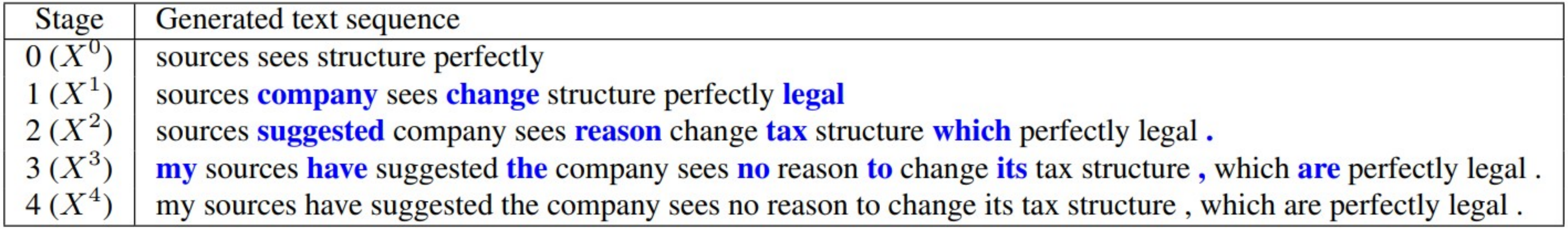
[NAR tutorial] Jiatao Gu, Xu Tan: Non-Autoregressive Sequence Generation
Tutorial at ACL 2022, May 22, 2022, https://github.com/NAR-tutorial/acl2022

IDA 2023

Method

Today’s focus: insertion models

Principle [Stern 2018]



Stage	Generated text sequence
0 (X^0)	sources sees structure perfectly
1 (X^1)	sources **company** sees **change** structure perfectly **legal**
2 (X^2)	sources **suggested** company sees **reason** change **tax** structure **which** perfectly legal **.**
3 (X^3)	**my** sources **have** suggested **the** company sees **no** reason **to** change **its** tax structure **,** which **are** perfectly legal .
4 (X^4)	my sources have suggested the company sees no reason to change its tax structure , which are perfectly legal .

Properties

→ can build sentence around key words

→ can guarantee tokens presence + order

→ can restrict where to fill with left an right context (unused here)

[Stern 2018] Mitchell Stern, William Chan, Jamie Kiros, and Jakob Uszkoreit:
Insertion Transformer: Flexible Sequence Generation via Insertion Operations
In: Proceedings of International Conference on Machine Learning. ICML (2018)

Method

Implementation

Survey → our own → POINTER [Zhang 2020]

(not to be confused with Pointer Networks [Vinyals 2015])

(+) use BERT skeleton, can leverage BERT checkpoints

(+) open source, give datasets and pretrained models

(-) no cross-attention

[Zhang 2020] Yizhe Zhang, Guoyin Wang, Chunyuan Li, Zhe Gan, Chris Brockett, Bill Dolan:
POINTER: Constrained progressive text generation via insertion-based generative pre-training.
In: Proceedings of EMNLP. pp. 8649–8670. ACL (2020)

[Vinyals 2015] Oriol Vinyals, Meire Fortunato, N. Jaitly:
Pointer Networks
In: In Advances in Neural Information Processing Systems, Vol. 28. NeurIPS (2015)

Method

Application: paraphrasing for data augmentation

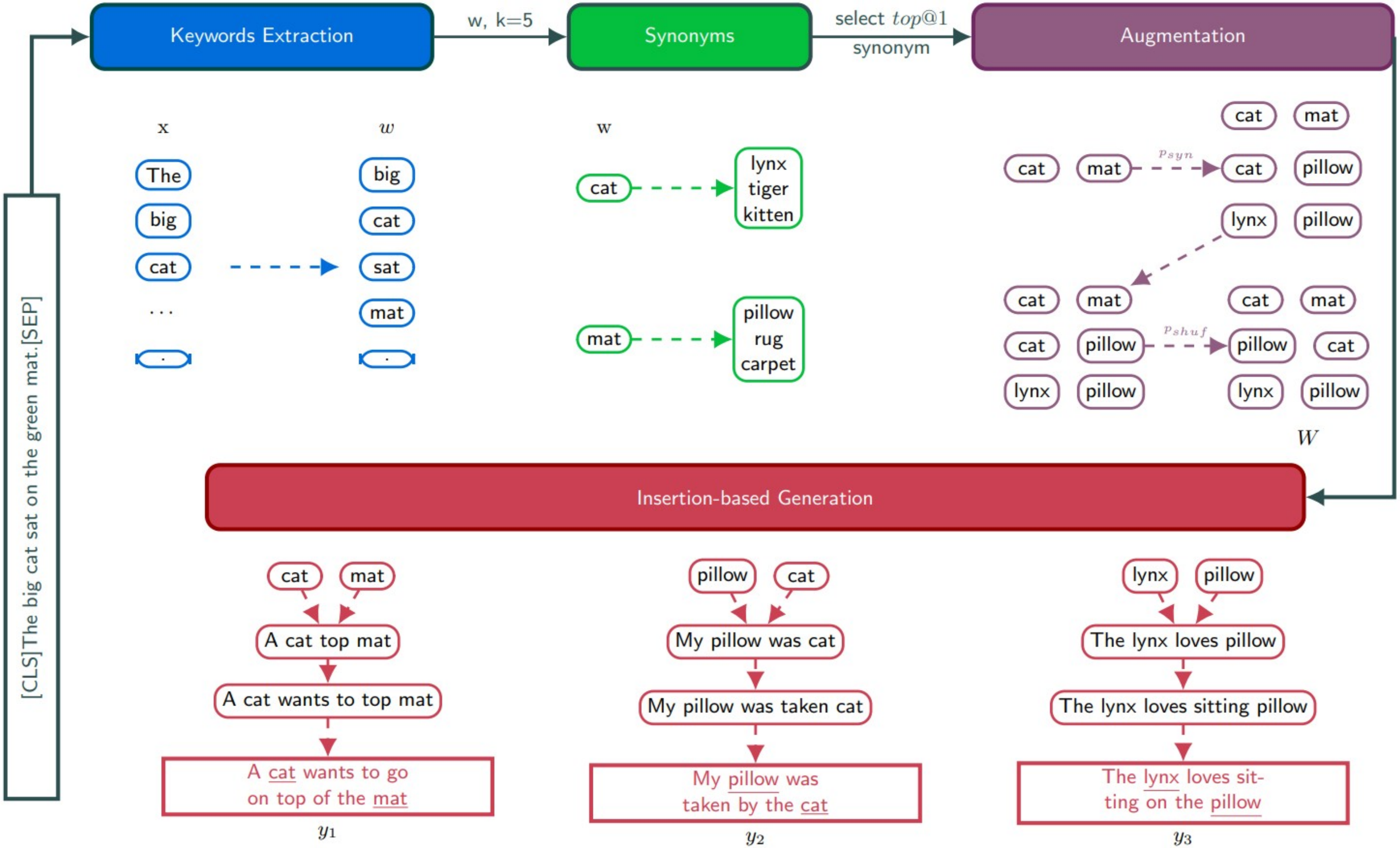
→ Why: high diversity, low fidelity

good match w/ no cross-attention

→ How: ...

Method - overview

→ How:



[CLS]The big cat sat on the green mat.[SEP]

Keywords Extraction

w, k=5

Synonyms

select $top@1$ synonym

Augmentation

x: The, big, cat, ..., .

w: big, cat, sat, mat, .

w: cat → lynx, tiger, kitten; mat → pillow, rug, carpet

cat mat

cat mat $\xrightarrow{p_{syn}}$ cat pillow

lynx pillow

cat mat — cat mat

cat pillow $\xrightarrow{p_{shuf}}$ pillow cat

lynx pillow — lynx pillow

W

Insertion-based Generation

cat, mat → A cat top mat → A cat wants to top mat → A cat wants to go on top of the mat

y_1

pillow, cat → My pillow was cat → My pillow was taken cat → My pillow was taken by the cat

y_2

lynx, pillow → The lynx loves pillow → The lynx loves sitting pillow → The lynx loves sitting on the pillow

y_3

Method - detail

Keyword extraction:

Identify the k (=5) most important keywords

a keyword = a list of token e.g. « every#day », « living room »

← trained keyword extractor – stopwords + end punctuation

Keyword augmentation:

- shuffle each keyword with proba p_{shuf} (=0,75)

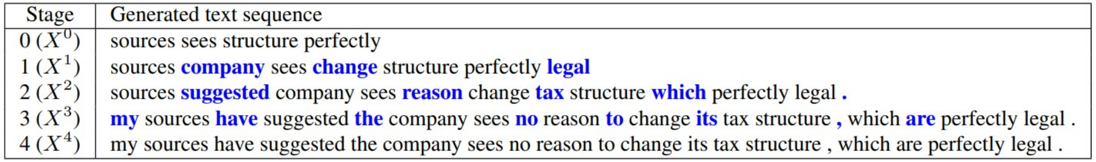
- synonimize each keyword with proba p_{syn} (=0,25)

→ gives m (=5) different augmentations

Constrained generation:

Finetuned POINTER model

...



Stage	Generated text sequence
0 (X^0)	sources sees structure perfectly
1 (X^1)	sources **company** sees **change** structure perfectly **legal**
2 (X^2)	sources **suggested** company sees **reason** change **tax** structure **which** perfectly legal **.**
3 (X^3)	**my** sources **have** suggested **the** company sees **no** reason **to** change **its** tax structure **,** which **are** perfectly legal .
4 (X^4)	my sources have suggested the company sees no reason to change its tax structure , which are perfectly legal .

Table 1: Example of the progressive generation process with multiple stages from the POINTER model. Words in **blue** indicate newly generated words at the current stage. X^i denotes the generated partial sentence at Stage i X^4 and X^3 are the same indicates the end of the generation process. Interestingly, our method allows informative words (*e.g.*, *company, change*) generated before the non-informative words (*e.g.*, *the, to*) generated at the end.

Metrics - separable

Evaluation

A batch of paraphrases should be:
1) fluent
2) diverse (from source + each others)
3) semantically similar

We can resp. use:
1) ppl (v)
2) dist-2 (^)
3) use-similarity (^)
as a quantitative approximation + qualitative look

Metrics - separable

Example with “A big cat”→ “A huge cat”:

1) ppl(“A huge cat”)

    [p(“A”) * p(“huge”) * p(“cat”)] ^ -1/3 = 11.3

2) dist-2(“A huge cat”, “A big cat”)

    4 distinct bi-grams / 4 distinct tokens = 1

3) use-similarity(“A huge cat”, “A big cat”)

    cos(embedding(“A huge cat”), embedding(“A big cat”)) = 0.943

Metrics - global

Extra metrics

1) BLEURT

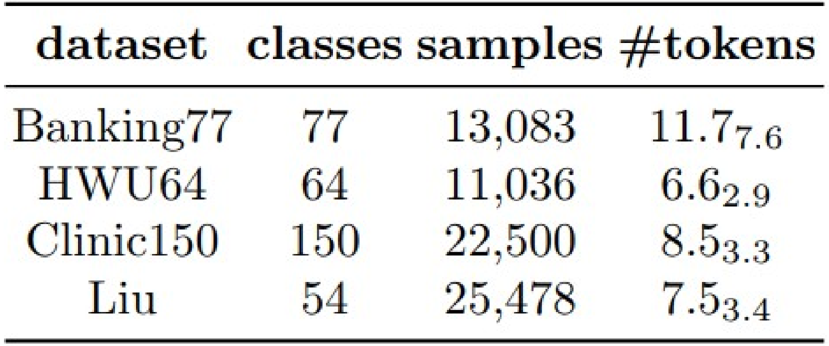
2) BERTScore

3) End task
   classification accuracy w/ data augmentation

Experiments

Datasets

4 fine-grained classification datasets



dataset	classes	samples	#tokens
Banking77	77	13,083	$11.7_{7.6}$
HWU64	64	11,036	$6.6_{2.9}$
Clinic150	150	22,500	$8.5_{3.3}$
Liu	54	25,478	$7.5_{3.4}$

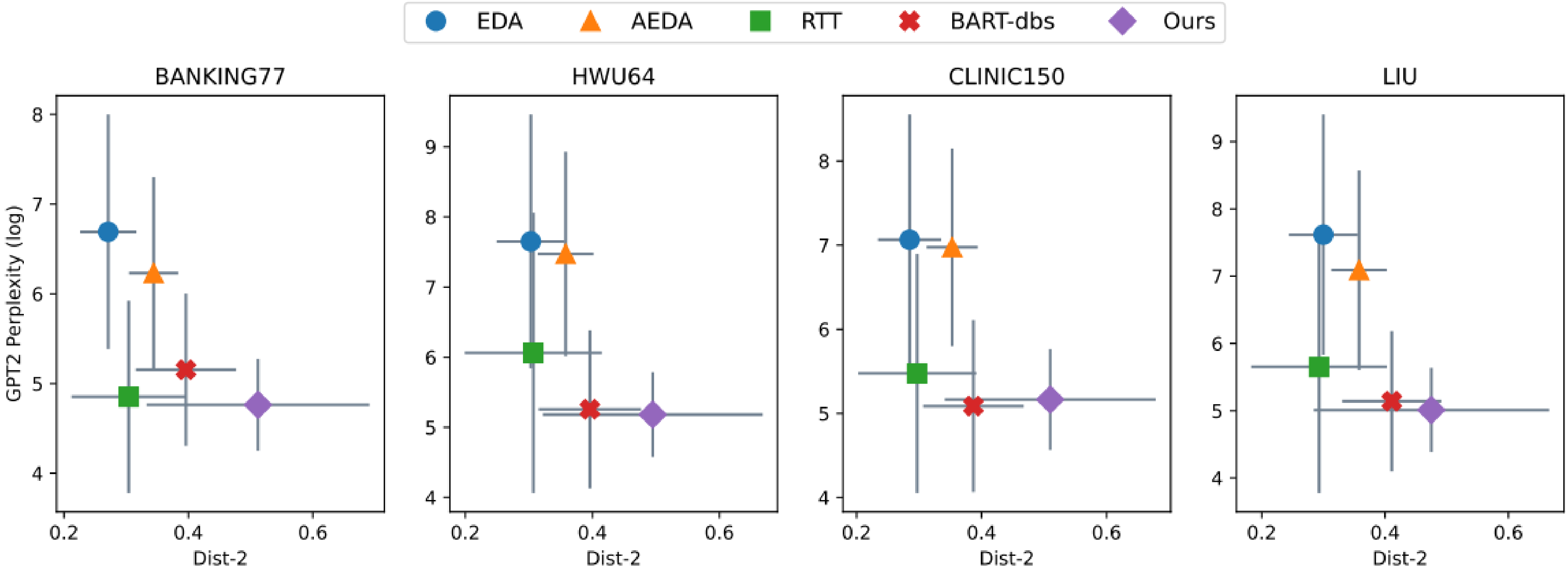
Baselines

- EDA (4 augmentations)
- AEDA (‘;’ insertion)
- RTT (⇌ translation)
- Bart-uni (finetuned BART with diverse beam search)

IDA 2023

Results - tradeoffs (separable metrics)

- dist-2 & use-similarity highly correlated → 2D tradeoff analysis
- we plot fluency (y) vs. diversity (x) for different baselines (color)
- best is fluent and diverse (lower-right corner) → ours (purple)



IDA 2023

Results - distribution (separable metrics)

- how are paraphrases distributed ?

orange: ours

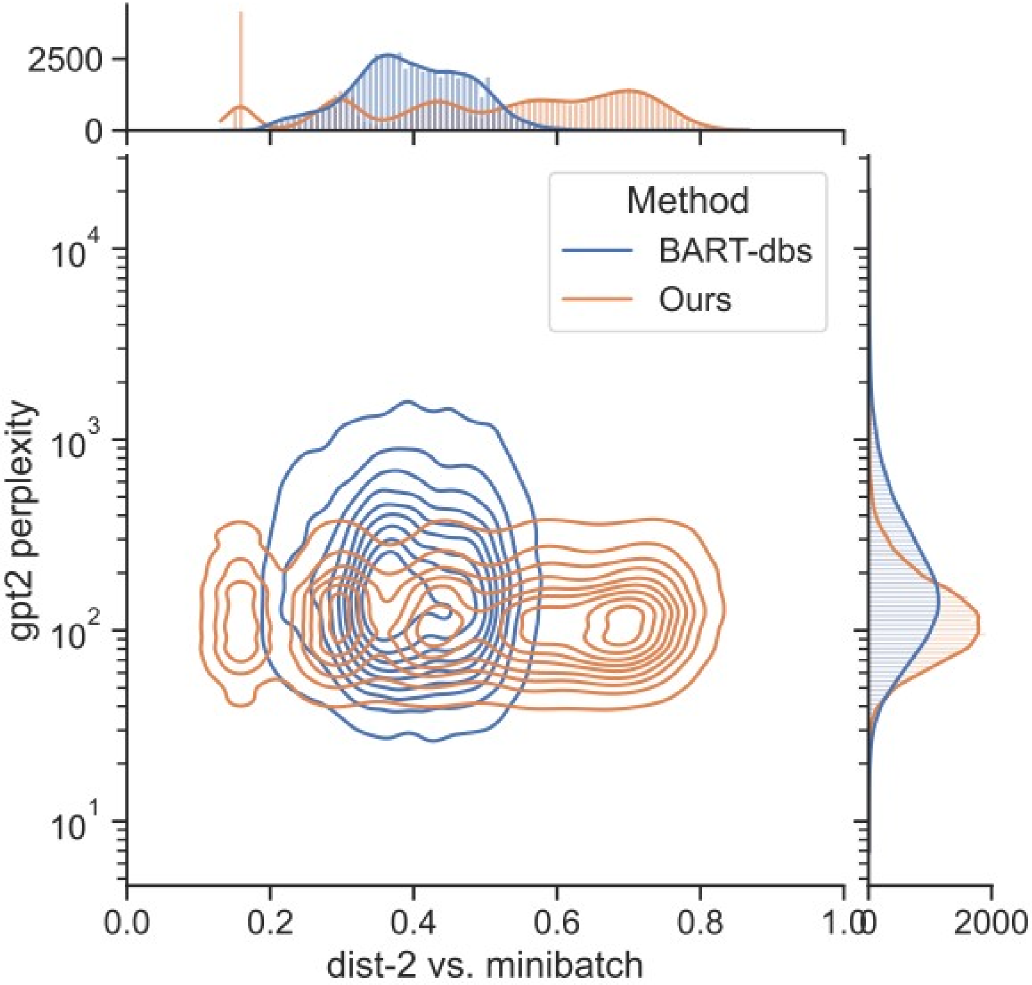
blue: best baseline

- less fluency spread

more diversity spread

- blobs ← number of augmented keywords

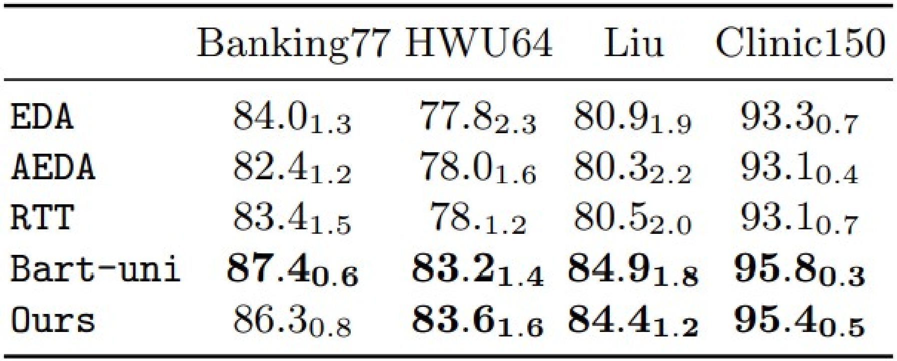
- fluency and diversity not correlated



Results - quantitative

→ Classification accuracy:

How do we perform as a data augmentation technique ?

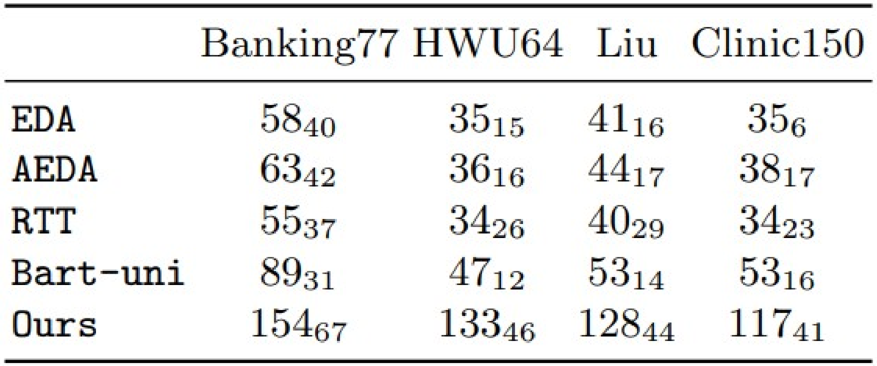


	Banking77	HWU64	Liu	Clinic150
EDA	$84.0_{1.3}$	$77.8_{2.3}$	$80.9_{1.9}$	$93.3_{0.7}$
AEDA	$82.4_{1.2}$	$78.0_{1.6}$	$80.3_{2.2}$	$93.1_{0.4}$
RTT	$83.4_{1.5}$	$78._{1.2}$	$80.5_{2.0}$	$93.1_{0.7}$
Bart-uni	$\mathbf{87.4_{0.6}}$	$\mathbf{83.2_{1.4}}$	$\mathbf{84.9_{1.8}}$	$\mathbf{95.8_{0.3}}$
Ours	$86.3_{0.8}$	$\mathbf{83.6_{1.6}}$	$\mathbf{84.4_{1.2}}$	$\mathbf{95.4_{0.5}}$

we match our best AR baseline

→ Length mismatch:

Table 2: Average number of characters of the paraphrases $mean_{std}$.



	Banking77	HWU64	Liu	Clinic150
EDA	58_{40}	35_{15}	41_{16}	35_{6}
AEDA	63_{42}	36_{16}	44_{17}	38_{17}
RTT	55_{37}	34_{26}	40_{29}	34_{23}
Bart-uni	89_{31}	47_{12}	53_{14}	53_{16}
Ours	154_{67}	133_{46}	128_{44}	117_{41}

- ⇌ translation: x0.96
- finetuned AR model: x1.4
- ours: x3.1

Results - qualitative

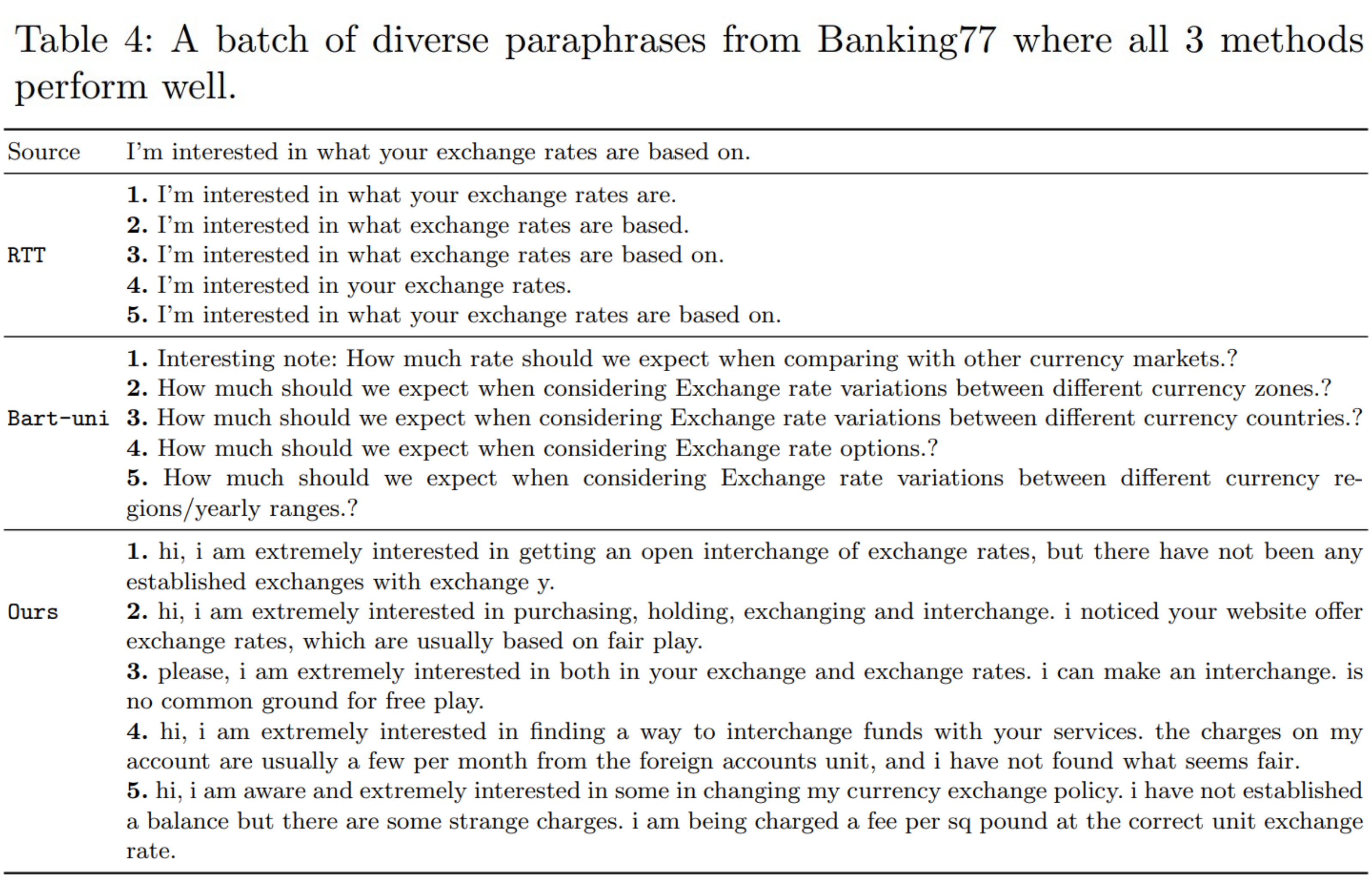


Table 4: A batch of diverse paraphrases from Banking77 where all 3 methods perform well.

Source	I'm interested in what your exchange rates are based on.
RTT	**1.** I'm interested in what your exchange rates are. **2.** I'm interested in what exchange rates are based. **3.** I'm interested in what exchange rates are based on. **4.** I'm interested in your exchange rates. **5.** I'm interested in what your exchange rates are based on.
Bart-uni	**1.** Interesting note: How much rate should we expect when comparing with other currency markets.? **2.** How much should we expect when considering Exchange rate variations between different currency zones.? **3.** How much should we expect when considering Exchange rate variations between different currency countries.? **4.** How much should we expect when considering Exchange rate options.? **5.** How much should we expect when considering Exchange rate variations between different currency regions/yearly ranges.?
Ours	**1.** hi, i am extremely interested in getting an open interchange of exchange rates, but there have not been any established exchanges with exchange y. **2.** hi, i am extremely interested in purchasing, holding, exchanging and interchange. i noticed your website offer exchange rates, which are usually based on fair play. **3.** please, i am extremely interested in both in your exchange and exchange rates. i can make an interchange. is no common ground for free play. **4.** hi, i am extremely interested in finding a way to interchange funds with your services. the charges on my account are usually a few per month from the foreign accounts unit, and i have not found what seems fair. **5.** hi, i am aware and extremely interested in some in changing my currency exchange policy. i have not established a balance but there are some strange charges. i am being charged a fee per sq pound at the correct unit exchange rate.

Results - qualitative

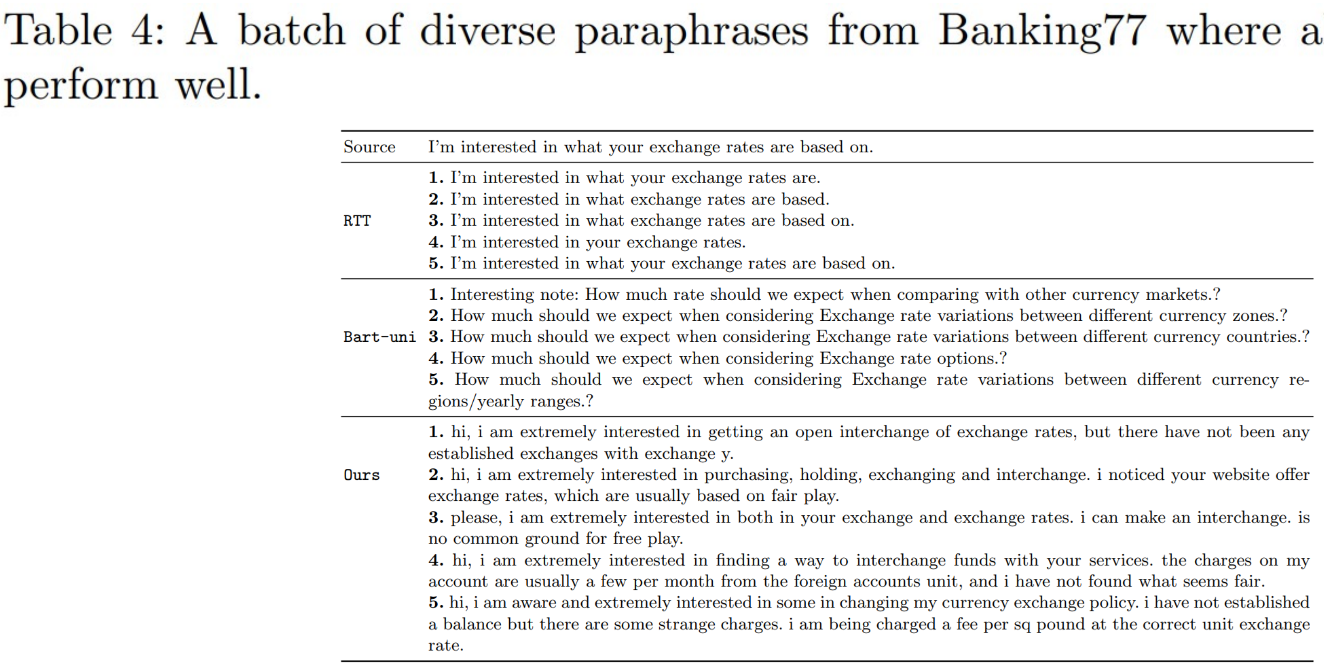


Table 4: A batch of diverse paraphrases from Banking77 where all 3 methods perform well.

Source	I'm interested in what your exchange rates are based on.
RTT	**1.** I'm interested in what your exchange rates are. **2.** I'm interested in what exchange rates are based. **3.** I'm interested in what exchange rates are based on. **4.** I'm interested in your exchange rates. **5.** I'm interested in what your exchange rates are based on.
Bart-uni	**1.** Interesting note: How much rate should we expect when comparing with other currency markets.? **2.** How much should we expect when considering Exchange rate variations between different currency zones.? **3.** How much should we expect when considering Exchange rate variations between different currency countries.? **4.** How much should we expect when considering Exchange rate options.? **5.** How much should we expect when considering Exchange rate variations between different currency regions/yearly ranges.?
Ours	**1.** hi, i am extremely interested in getting an open interchange of exchange rates, but there have not been any established exchanges with exchange y. **2.** hi, i am extremely interested in purchasing, holding, exchanging and interchange. i noticed your website offer exchange rates, which are usually based on fair play. **3.** please, i am extremely interested in both in your exchange and exchange rates. i can make an interchange. is no common ground for free play. **4.** hi, i am extremely interested in finding a way to interchange funds with your services. the charges on my account are usually a few per month from the foreign accounts unit, and i have not found what seems fair. **5.** hi, i am aware and extremely interested in some in changing my currency exchange policy. i have not established a balance but there are some strange charges. i am being charged a fee per sq pound at the correct unit exchange rate.

- aug/⇌T: conservative - minor edits
- seq2seq: more creative - start to add some info, still see where the beam search (// gen process) diverge
- from keywords: very creative, for better and for worse

Perspectives

Limitations

→ involved pipeline
→ length mismatch } → can opt for an end-to-end cross-attention model
→ or can control with a latent/input variable

Strengths

→ very extensible (+ more interpretable) generation

Conclusion

→ insertion models offers more flexibility and control, but harder to train
→ NAR is viable for data augmentation on low-resource fine classif
→ exciting future for semi-AR and NAR methods
open new tasks + can benefit e.g. interactive writing
(change style/length of a span w/ context and better Pareto)

References

[Yang 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le:
XLNet: Generalized Autoregressive Pretraining for Language Understanding
In: Proceedings of Advances in Neural Information Processing Systems 32. NeurIPS (2019)

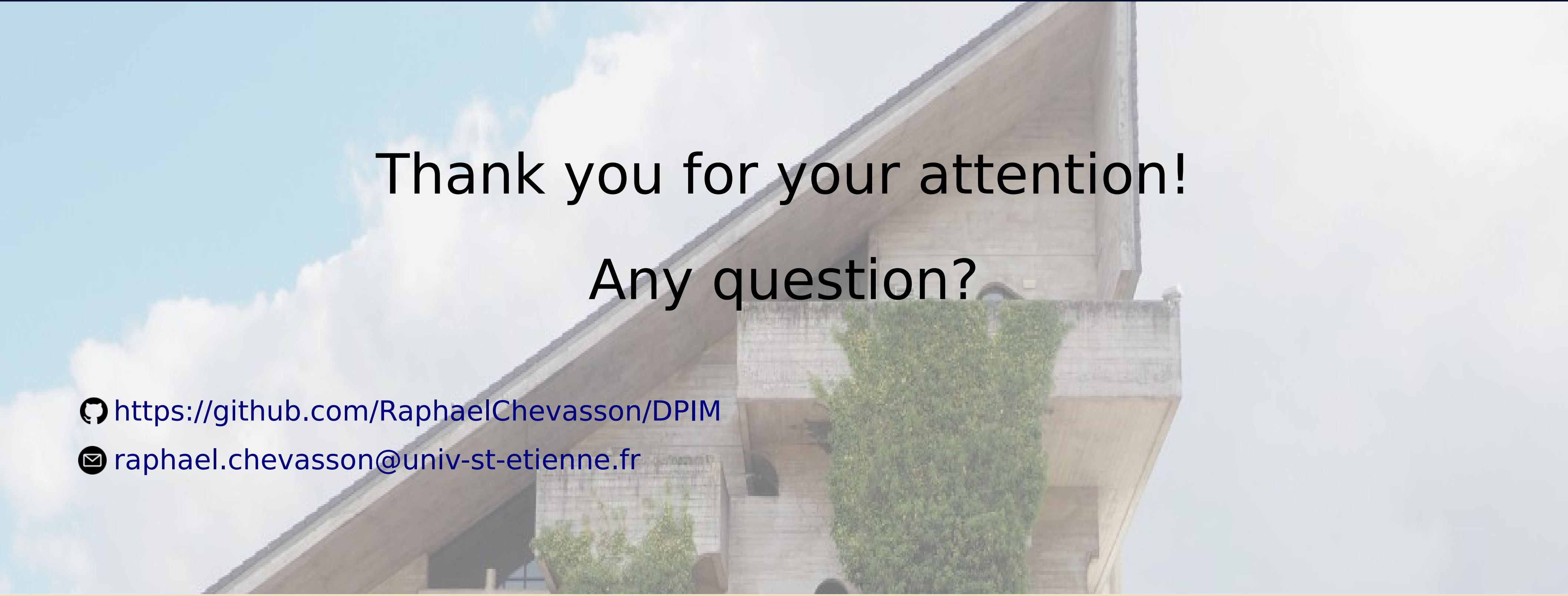
[Gu 2017] Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, Richard Socher:
Non-Autoregressive Neural Machine Translation
In: Proceedings of International Conference on Learning Representations. ICLR (2018)

[Gu 2022] Jiatao Gu, Xu Tan:
Non-Autoregressive Sequence Generation
Tutorial at ACL 2022, May 22, 2022, https://github.com/NAR-tutorial/acl2022

[Stern 2018] Mitchell Stern, William Chan, Jamie Kiros, and Jakob Uszkoreit:
Insertion Transformer: Flexible Sequence Generation via Insertion Operations
In: Proceedings of International Conference on Machine Learning. ICML (2018)

[Zhang 2020] Yizhe Zhang, Guoyin Wang, Chunyuan Li, Zhe Gan, Chris Brockett, Bill Dolan:
POINTER: Constrained progressive text generation via insertion-based generative pre-training.
In: Proceedings of EMNLP. pp. 8649–8670. ACL (2020)

[Vinyals 2015] Oriol Vinyals, Meire Fortunato, N. Jaitly:
Pointer Networks
In: In Advances in Neural Information Processing Systems, Vol. 28. NeurIPS (2015)



Thank you for your attention!

Any question?

https://github.com/RaphaelChevasson/DPIM

raphael.chevasson@univ-st-etienne.fr

IDA 2023

Results - quantitative

→ Additional results – overall quality:

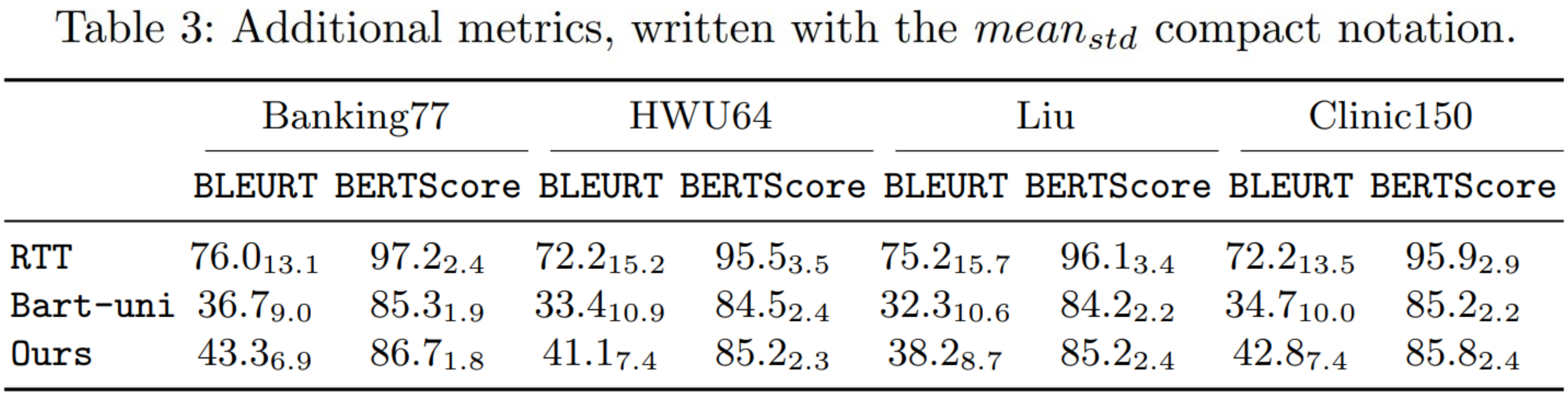


Table 3: Additional metrics, written with the $mean_{std}$ compact notation.

	Banking77		HWU64		Liu		Clinic150	
	BLEURT	BERTScore	BLEURT	BERTScore	BLEURT	BERTScore	BLEURT	BERTScore
RTT	$76.0_{13.1}$	$97.2_{2.4}$	$72.2_{15.2}$	$95.5_{3.5}$	$75.2_{15.7}$	$96.1_{3.4}$	$72.2_{13.5}$	$95.9_{2.9}$
Bart-uni	$36.7_{9.0}$	$85.3_{1.9}$	$33.4_{10.9}$	$84.5_{2.4}$	$32.3_{10.6}$	$84.2_{2.2}$	$34.7_{10.0}$	$85.2_{2.2}$
Ours	$43.3_{6.9}$	$86.7_{1.8}$	$41.1_{7.4}$	$85.2_{2.3}$	$38.2_{8.7}$	$85.2_{2.4}$	$42.8_{7.4}$	$85.8_{2.4}$