Measuring Immediate Adaptation Performance for Neural Machine Translation

Patrick Simianer, Joern Wuebker, John DeNero

Lilt

NAACL '19

Outline

1. Motivation & Approach
2. Evaluation
3. Conclusion

Motivation

Online adaptation is a key feature of modern computer-aided translation (CAT)

Motivation

Online adaptation is a key feature of modern computer-aided translation (CAT)

Non-adaptive system

Source #1: Der Terrier beißt die Frau

Motivation

Online adaptation is a key feature of modern computer-aided translation (CAT)

Non-adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady

Motivation

Online adaptation is a key feature of modern computer-aided translation (CAT)

Non-adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier bites the woman

Motivation

Online adaptation is a key feature of modern computer-aided translation (CAT)

Non-adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier bites the woman
Source #2:	Der Mann beißt den Terrier

Motivation

Online adaptation is a key feature of modern computer-aided translation (CAT)

Non-adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier bites the woman
Source #2:	Der Mann beißt den Terrier
Hypothesis #2:	The dog bites the man

Motivation

Online adaptation is a key feature of modern computer-aided translation (CAT)

Non-adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier bites the woman
Source #2:	Der Mann beißt den Terrier
Hypothesis #2:	The dog bites the man
Reference #2:	The man bites the terrier

Motivation

Translators have a reasonable expectation that . . .

1. **New vocabulary** (in context) gets quickly picked up by the system, ideally right away
2. The system generally adapts to **new domains**

Motivation

Translators have a reasonable expectation that . . .

1. **New vocabulary** (in context) gets quickly picked up by the system, ideally right away
2. The system generally adapts to **new domains**

With **neural machine translation** *fine-tuning* can readily be used [Turchi et al., 2017] (*inter-alia*):

$$\theta_i \leftarrow \theta_{i-1} - \gamma \nabla \mathcal{L}(\theta_{i-1}, x_i, y_i).$$

Approach

- Typically [Turchi et al., 2017, Peris et al., 2017, Bertoldi et al., 2014] (*inter-alia*) fine-tuning is evaluated in a batch setting
- Corpus BLEU or isolated sentence-wise metrics are often used
- These do not necessarily express how fast a system adapts

Approach

- Typically [Turchi et al., 2017, Peris et al., 2017, Bertoldi et al., 2014] (*inter-alia*) fine-tuning is evaluated in a batch setting
- Corpus BLEU or isolated sentence-wise metrics are often used
- These do not necessarily express how fast a system adapts

As we will show this is not good enough

→ We seek to measure perceived, **immediate** adaptation performance

Approach

Calculate **recall** on the set of all words that are not stopwords, ignoring length [Papineni et al., 2002] and ordering issues[1] [Kothur et al., 2018]

[1] In each of the data sets considered in this work, the average number of occurrences of content words ranges between 1.01 and 1.11 per sentence

Approach

Calculate **recall** on the set of all words that are not stopwords, ignoring length [Papineni et al., 2002] and ordering issues[1] [Kothur et al., 2018]

Since the task is online adaptation — specifically focus on **few-shot learning**: Consider only **first** and **second** occurrences of words!

[1] In each of the data sets considered in this work, the average number of occurrences of content words ranges between 1.01 and 1.11 per sentence

One-Shot Recall R1

After seeing a word exactly once before in a reference/confirmed translation, is it correctly produced the second time around?

One-Shot Recall R1

After seeing a word exactly once before in a reference/confirmed translation, is it correctly produced the second time around?

$$\mathrm{R1}_i = \frac{|\mathcal{H}_i \cap \mathcal{R}_{1,i}|}{|\mathcal{R}_{1,i}|}$$

$\mathcal{H}_i$: Content words in the hypothesis *i*th example

$\mathcal{R}_{1,i}$: Content words whose **second occurrence** is in the reference for *i*th example

One-Shot Recall R1: Example

Adaptive system

Source #1: Der Terrier beißt die Frau

One-Shot Recall R1: Example

Adaptive system

Source #1: Der Terrier beißt die Frau

Hypothesis #1: The dog bites the lady

One-Shot Recall R1: Example

Adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier bites the woman

One-Shot Recall R1: Example

Adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier bites the woman
	R1=0/0

One-Shot Recall R1: Example

Adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier bites the woman
	R1=0/0
Source #2:	Der Mann beißt den Terrier

One-Shot Recall R1: Example

Adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier bites the woman
	R1=0/0
Source #2:	Der Mann beißt den Terrier
Hypothesis #2:	The terrier bites the man

One-Shot Recall R1: Example

Adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier bites the woman
	R1=0/0
Source #2:	Der Mann beißt den Terrier
Hypothesis #2:	The terrier bites the man
Reference #2:	The man bites_1 the terrier_1

One-Shot Recall R1: Example

Adaptive system

Source #1: Der Terrier beißt die Frau

Hypothesis #1: The dog bites the lady

Reference #1: The terrier bites the woman

R1=0/0

Source #2: Der Mann beißt den Terrier

Hypothesis #2: The terrier bites the man

Reference #2: The man bites$_1$ the terrier$_1$

R1=2/2

One-Shot Recall R1: Example

Adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier bites the woman
	R1=0/0
Source #2:	Der Mann beißt den Terrier
Hypothesis #2:	The terrier bites the man
Reference #2:	The man bites$_1$ the terrier$_1$
	R1=$2/2$
Total:	R1=$2/2$

Zero-Shot Recall R0

Not having seen a word before, is it still correctly produced? Is the system adapting to the domain at hand?

Zero-Shot Recall R0

Not having seen a word before, is it still correctly produced? Is the system adapting to the domain at hand?

$$\mathrm{R0}_i = \frac{|\mathcal{H}_i \cap \mathcal{R}_{0,i}|}{|\mathcal{R}_{0,i}|}$$

$\mathcal{H}_i$: Content words in the hypothesis for *i*th example

$\mathcal{R}_{0,i}$: Content words that occur for the **first time** in the reference for *i*th example

Zero- and One-Shot Recall R0+1

Combined metric.

$$\text{R0+1}_i = \frac{|\mathcal{H}_i \cap [\mathcal{R}_{0,i} \cup \mathcal{R}_{1,i}]\,|}{|\mathcal{R}_{0,i} \cup \mathcal{R}_{1,i}|}$$

$\mathcal{H}_i$: Content words in the hypothesis for *i*th example

$\mathcal{R}_{0,i} \cup \mathcal{R}_{1,i}$: Content words that occur for the **first or second time** in the reference for *i*th example

Corpus-Level Metric

$$R0_{\text{Corpus}} = \frac{\sum_{i=1}^{|\mathcal{G}|} |\mathcal{H}_i \cap \mathcal{R}_{0,i}|}{\sum_{i=1}^{|\mathcal{G}|} |\mathcal{R}_{0,i}|}$$

$\mathcal{G}$: Corpus of $|\mathcal{G}|$ source, reference/confirmed segment, hypothesis triplets

Complete Example

Adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier$_0$ bites$_0$ the woman$_0$
	R1=$0/0$

Complete Example

Adaptive system

Source #1: Der Terrier beißt die Frau

Hypothesis #1: The dog bites the lady

Reference #1: The terrier$_0$ bites$_0$ the woman$_0$

R1=0/0 R0=1/3

Complete Example

Adaptive system

Source #1: Der Terrier beißt die Frau

Hypothesis #1: The dog bites the lady

Reference #1: The terrier$_0$ bites$_0$ the woman$_0$

$R1=0/0$ $R0=1/3$ $R0+1=1/3$

Complete Example

Adaptive system

Source #1: Der Terrier beißt die Frau

Hypothesis #1: The dog bites the lady

Reference #1: The terrier$_0$ bites$_0$ the woman$_0$

R1=0/0   R0=1/3   R0+1=1/3

Source #2: Der Mann beißt den Terrier

Complete Example

Adaptive system

Source #1: Der Terrier beißt die Frau

Hypothesis #1: The dog bites the lady

Reference #1: The terrier_0 bites_0 the woman_0

$R1=0/0$ $R0=1/3$ $R0+1=1/3$

Source #2: Der Mann beißt den Terrier

Hypothesis #2: The terrier bites the man

Complete Example

Adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier$_0$ bites$_0$ the woman$_0$
	R1=0/0 R0=1/3 R0+1=1/3
Source #2:	Der Mann beißt den Terrier
Hypothesis #2:	The terrier bites the man
Reference #2:	The man$_0$ bites$_1$ the terrier$_1$

Complete Example

Adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier$_0$ bites$_0$ the woman$_0$
	$R1=0/0$ $R0=1/3$ $R0+1=1/3$
Source #2:	Der Mann beißt den Terrier
Hypothesis #2:	The terrier bites the man
Reference #2:	The man$_0$ bites$_1$ the terrier$_1$
	$R1=2/2$

Complete Example

Adaptive system

Source #1: Der Terrier beißt die Frau

Hypothesis #1: The dog bites the lady

Reference #1: The terrier$_0$ bites$_0$ the woman$_0$

R1=0/0 R0=1/3 R0+1=1/3

Source #2: Der Mann beißt den Terrier

Hypothesis #2: The terrier bites the man

Reference #2: The man$_0$ bites$_1$ the terrier$_1$

R1=2/2 R0=1/1

Complete Example

Adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier$_0$ bites$_0$ the woman$_0$
	R1=$0/0$ R0=$1/3$ R0+1=$1/3$

Source #2:	Der Mann beißt den Terrier
Hypothesis #2:	The terrier bites the man
Reference #2:	The man$_0$ bites$_1$ the terrier$_1$
	R1=$2/2$ R0=$1/1$ R0+1=$3/3$

Complete Example

Adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier$_0$ bites$_0$ the woman$_0$
	R1=0/0 R0=1/3 R0+1=1/3
Source #2:	Der Mann beißt den Terrier
Hypothesis #2:	The terrier bites the man
Reference #2:	The man$_0$ bites$_1$ the terrier$_1$
	R1=2/2 R0=1/1 R0+1=3/3
Totals:	R1=2/2

Complete Example

Adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier$_0$ bites$_0$ the woman$_0$
	R1=0/0 R0=1/3 R0+1=1/3
Source #2:	Der Mann beißt den Terrier
Hypothesis #2:	The terrier bites the man
Reference #2:	The man$_0$ bites$_1$ the terrier$_1$
	R1=2/2 R0=1/1 R0+1=3/3
Totals:	R1=2/2 R0=2/4

Complete Example

Adaptive system

Source #1:	Der Terrier beißt die Frau
Hypothesis #1:	The dog bites the lady
Reference #1:	The terrier$_0$ bites$_0$ the woman$_0$
	R1=0/0 R0=1/3 R0+1=1/3
Source #2:	Der Mann beißt den Terrier
Hypothesis #2:	The terrier bites the man
Reference #2:	The man$_0$ bites$_1$ the terrier$_1$
	R1=2/2 R0=1/1 R0+1=3/3
Totals:	R1=2/2 R0=2/4 R0+1=4/6

Evaluation: Adaptation Methods

The task is **online adaptation** to the *Autodesk* data set [Zhechev, 2012]**.** The **background model** is an English-to-German Transformer, trained on about 100M segments.

Evaluation: Adaptation Methods

The task is **online adaptation** to the *Autodesk* data set [Zhechev, 2012]. The **background model** is an English-to-German Transformer, trained on about 100M segments.

Four methods for comparison:

bias Add an additional bias to the output projection [Michel and Neubig, 2018]

full Fine-tuning of all weights

top Adapt top encoder/decoder layers only

lasso Dynamic selection of adapted tensors with group lasso regularization [Wuebker et al., 2018]

Results

Results contrasting traditional MT metrics — BLEU, and TER — to the proposed metrics.

Relative differences for adaptive systems, positive results highlighted with green color.

System ↓ / Metric →	BLEU	TER	R1	R0	R0+1
baseline	40.3	45.2	44.9	39.3	41.0
bias	0	0	1	0	0
full	17	-3	22	-9	1
top	7	10	12	-9	-2
lasso	15	-6	8	3	4

Results: Novel Content Words

Results when calculating the metrics only for truly novel content words, i.e. ones that do not occur in the training data.

System ↓ / Metric →	R1	R0	R0+1
baseline	27.1	40.7	29.9
full	**55**	-4	13
lasso	30	**18**	**21**

Conclusion

- **Immediate adaptation performance** is important for adaptive MT in CAT
- We proposed **three metrics** for measuring immediate and possibly perceived adaptation performance
  - R1 for **one-shot recall**, quantifying pick up of new vocabulary
  - R0 for **zero-shot recall**, quantifying general domain adaptation performance
  - The combined metric R0+1
- These metrics give a **different signal** than the MT metrics that are traditionally used
- Zero-shot recall **R0 suffers from unregularized adaptation**!
- **Careful regularization** can mitigate this effect, while retaining most of the one-shot recall R1

Conclusion

- **Immediate adaptation performance** is important for adaptive MT in CAT
- We proposed **three metrics** for measuring immediate and possibly perceived adaptation performance
    - R1 for **one-shot recall**, quantifying pick up of new vocabulary
    - R0 for **zero-shot recall**, quantifying general domain adaptation performance
    - The combined metric R0+1
- These metrics give a **different signal** than the MT metrics that are traditionally used
- Zero-shot recall **R0 suffers from unregularized adaptation**!
- **Careful regularization** can mitigate this effect, while retaining most of the one-shot recall R1

Thank you!

Bibliography I

N. Bertoldi, P. Simianer, M. Cettolo, K. Wäschle, M. Federico, and S. Riezler. Online adaptation to post-edits for phrase-based statistical machine translation. *Machine Translation*, 28(3-4):309–339, 2014.

S. S. R. Kothur, R. Knowles, and P. Koehn. Document-level adaptation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 64–73, 2018.

P. Michel and G. Neubig. Extreme adaptation for personalized neural machine translation. *arXiv preprint arXiv:1805.01817*, 2018.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

Á. Peris, L. Cebrián, and F. Casacuberta. Online learning for neural machine translation post-editing. *arXiv preprint arXiv:1706.03196*, 2017.

Bibliography II

M. Turchi, M. Negri, M. A. Farajian, and M. Federico. Continuous learning from human post-edits for neural machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):233–244, 2017.

J. Wuebker, P. Simianer, and J. DeNero. Compact personalized models for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

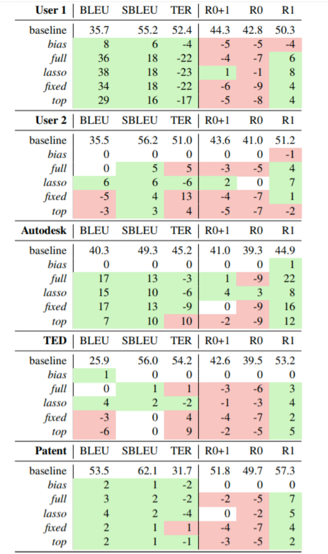
V. Zhechev. Machine translation infrastructure and post-editing performance at autodesk. In *AMTA 2012 workshop on post-editing technology and practice (WPTP 2012)*, pages 87–96. San Diego USA, 2012.

Results: Subwords

Results when calculating the metrics with subwords.

System ↓ / Metric →	R1	R0	R0+1
baseline	48.1	**44.1**	45.5
full	**14**	-8	0
lasso	7	-1	**2**

Complete Results Table



User 1	BLEU	SBLEU	TER	R0+1	R0	R1
baseline	35.7	55.2	52.4	44.3	42.8	50.3
bias	8	6	-4	-5	-5	-4
full	36	18	-22	-4	-7	6
lasso	38	18	-23	1	-1	8
fixed	34	18	-22	-6	-9	4
top	29	16	-17	-5	-8	4
User 2	BLEU	SBLEU	TER	R0+1	R0	R1
baseline	35.5	56.2	51.0	43.6	41.0	51.2
bias	0	0	0	0	0	-1
full	0	5	5	-3	-5	4
lasso	6	6	-6	2	0	7
fixed	-5	4	13	-4	-7	1
top	-3	3	4	-5	-7	-2
Autodesk	BLEU	SBLEU	TER	R0+1	R0	R1
baseline	40.3	49.3	45.2	41.0	39.3	44.9
bias	0	0	0	0	0	1
full	17	13	-3	1	-9	22
lasso	15	10	-6	4	3	8
fixed	17	13	-9	0	-9	16
top	7	10	10	-2	-9	12
TED	BLEU	SBLEU	TER	R0+1	R0	R1
baseline	25.9	56.0	54.2	42.6	39.5	53.2
bias	1	0	0	0	0	0
full	0	1	1	-3	-6	3
lasso	4	2	-2	-1	-3	4
fixed	-3	0	4	-4	-7	2
top	-6	0	9	-2	-5	5
Patent	BLEU	SBLEU	TER	R0+1	R0	R1
baseline	53.5	62.1	31.7	51.8	49.7	57.3
bias	2	1	-2	0	0	0
full	3	2	-2	-2	-5	7
lasso	4	2	-4	0	-2	5
fixed	2	1	1	-4	-7	4
top	2	1	-1	-3	-5	2