Document-level Machine Translation: Recent Progress and The Crux of Evaluation

Rico Sennrich

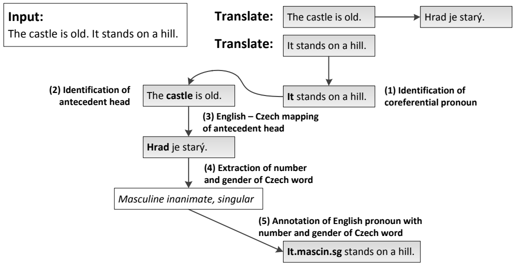
University of Zurich[UZH]

Edinburgh NLP
University of Edinburgh
Natural Language Processing

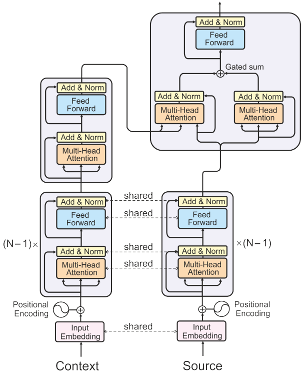
Setting the Scene: Why Document-level MT?

Microsoft reaches a historic milestone, using AI to match human performance in translating news from Chinese to English

March 14, 2018 | Allison Linn



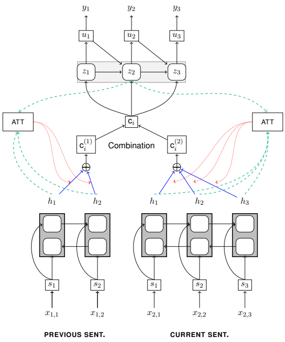
context-aware SMT architecture



context-aware NMT architecture

Setting the Scene: Multi-Source Architectures

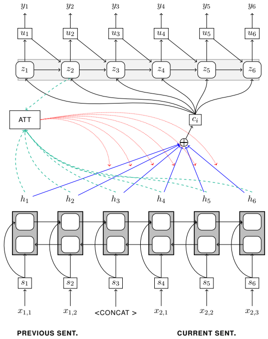
idea: use additional encoders for context [Jean et al., 2017, Wang et al., 2017]



[Bawden et al., 2018]

Setting the Scene: Concatenation Strategy

main translation unit: concatenation of multiple sentences



[Tiedemann and Scherrer, 2017]

[Bawden et al., 2018]

[Junczys-Dowmunt, 2019]: sequences of up to 1000 (sub)words
$\rightarrow$ enough to translate many news articles as one sequence.

Some Open Questions

- How do we measure progress?
- Which context matters?
- What neural architectures work well?
- How do we make sure model learns to consider context?
- How do we deal with lack of document-level data?

Some Open Questions

- **How do we measure progress?**
- Which context matters?
- What neural architectures work well?
- How do we make sure model learns to consider context?
- **How do we deal with lack of document-level data?**

Standard Metrics

problems with BLEU and other standard metrics:

- n-gram statistics are very local
  $\rightarrow$ insensitive to long-distance agreement etc.
- reference-based evaluation not ideal to measure consistency
  $\rightarrow$ consistency does not increase expected overlap with reference
- only small proportion of words depends on context beyond sentence
  $\rightarrow$ how do we measure incremental improvements?

Reference-Based Evaluation of Pronoun Translation

idea: let's automatically measure if translation of pronouns matches reference (APT; AutoPRF)
problem: this agrees poorly with human judgments ☹

example [Guillou and Hardmeier, 2018]

SOURCE: so what these two **clips** show is not just the devastating consequence of the disease, but **they** also tell us something about the shocking pace of the disease...

MT: donc ce que ces deux **extraits[masc.pl.]** montrent n'est pas seulement la consequence devastatrice de la maladie, mais **ils[masc.pl.]** nous disent aussi quelque chose sur le rythme choquant de la maladie...

REFERENCE: ce que ces deux **videos[fem.pl.]** montrent, ce ne sont pas seulement les consequences dramatiques de cette maladie, **elles[fem.pl.]** nous montrent aussi la vitesse fulgurante de cette maladie...

Repetition Rate as Cohesion Metric?

[Wong and Kit, 2012]: more cohesive translations have more repetitions

$$RC = \frac{\text{number of repeated words}}{\text{number of content words}}$$

Repetition Rate as Cohesion Metric?

problem:

sentence-level MT is (accidentally) more repetitive than human translation!

an artifact of statistical language modeling?

GPT-2-produced text

human-produced text

Hendrik Strobelt and Sebastian Gehrmann: http://gltr.io/

can we distinguish accidental repetition from document-level cohesion?

Evaluating Discourse Phenomena

[Bawden et al., NAACL 2018]




- targeted evaluation:
  hand-crafted test set of 200 context-dependent translations
- exploration of multi-encoder and concatenation architectures
- models trained on subset of OpenSubtitles2016 English-French

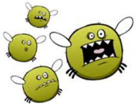
A Contrastive Test Set: Coreference

Source:

context:	Oh, I hate **flies**. Look, there's another one!
sentence:	Don't worry, I'll kill **it** for you.

Target:

context:	Ô je déteste les mouches. Regarde, il y en a une autre !
correct:	T'inquiète, je **la** tuerai pour toi.
incorrect:	T'inquiète, je **le** tuerai pour toi.



A Contrastive Test Set: Coreference

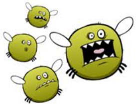
Can the model rank the correct sentence above the incorrect one?

Source:

context: Oh, I hate **flies**. Look, there's another one!
sentence: Don't worry, I'll kill **it** for you.

Target:

context: Ô je déteste les mouches.
Regarde, il y en a une autre !
correct: T'inquiète, je **la** tuerai pour toi.
incorrect: T'inquiète, je **le** tuerai pour toi.



A Contrastive Test Set: Coreference

Can the model rank the correct sentence above the incorrect one?

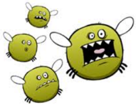
Previous linguistic context necessary to disambiguate

Source:

context: Oh, I hate **flies**. Look, there's another one!
sentence: Don't worry, I'll kill **it** for you.

Target:

context: Ô je déteste les mouches.
Regarde, il y en a une autre !
correct: T'inquiète, je **la** tuerai pour toi.
incorrect: T'inquiète, je **le** tuerai pour toi.



A Contrastive Test Set: Coreference

Can the model rank the correct sentence above the incorrect one?

Previous linguistic context necessary to disambiguate

Source:

context: Oh, I hate **flies**. Look, there's another one!

sentence: Don't worry, I'll kill **it** for you.

Target:

context: Ô je déteste les mouches. Regarde, il y en a une autre !

correct: T'inquiète, je **la** tuerai pour toi.

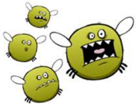
incorrect: T'inquiète, je **le** tuerai pour toi.

context: Ô je déteste les **moucherons**. Regarde, il y en a un autre !

correct: T'inquiète, je **le** tuerai pour toi.

incorrect: T'inquiète, je **la** tuerai pour toi.

Balanced examples: Non-contextual baseline scores 50%



A Contrastive Test Set: Coherence and Cohesion

Source:

context:	So what do you say to £50?
current sent.:	It's a little **steeper** than I was expecting.

Target:

context:	Qu'est-ce que vous en pensez de 50£ ?
correct:	C'est un peu plus **cher** que ce que je pensais.
incorrect:	C'est un peu plus **raide** que ce que je pensais.

Source:

context:	How are your feet holding up?
current sent.:	It's a little **steeper** than I was expecting.

Target:

context:	Comment vont tes pieds ?
correct:	C'est un peu plus **raide** que ce que je pensais.
incorrect:	C'est un peu plus **cher** que ce que je pensais.

A Contrastive Test Set: Coherence and Cohesion

Source:

context:	What's **crazy** about me?
current sent.:	Is this **crazy**?

Target:

context:	Qu'est-ce qu'il y a de **dingue** chez moi ?
correct:	Est-ce que ça c'est **dingue** ?
incorrect:	Est-ce que ça c'est fou ?

Source:

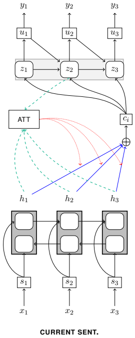
context:	What's **crazy** about me?
current sent.:	Is this **crazy**?

Target:

context:	Qu'est-ce qu'il y a de **fou** chez moi ?
correct:	Est-ce que ça c'est **fou** ?
incorrect:	Est-ce que ça c'est dingue ?

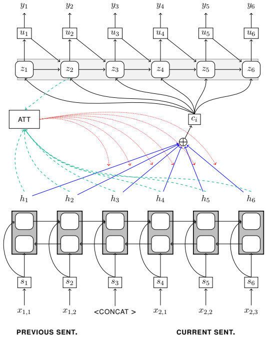
Case Study: Architectures

Baseline



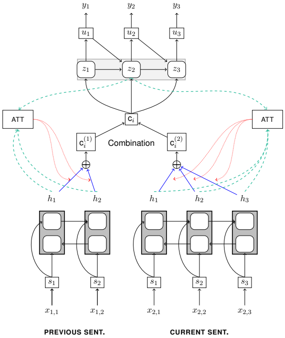
[Bahdanau et al., 2015]

2TO2 - concatenated input



[Tiedemann and Scherrer, 2017]

Multiple encoders



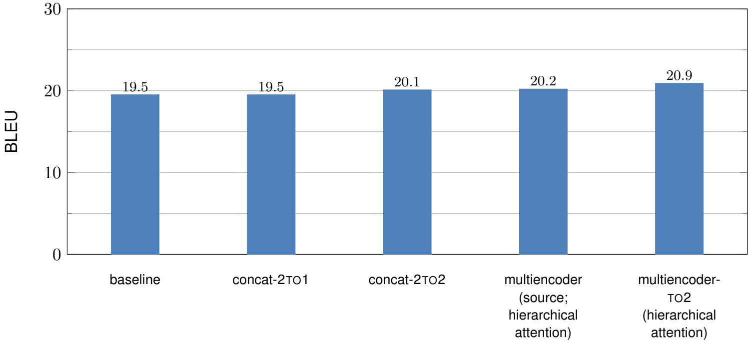
[Jean et al., 2017, Wang et al., 2017]

Case Study: Architectures

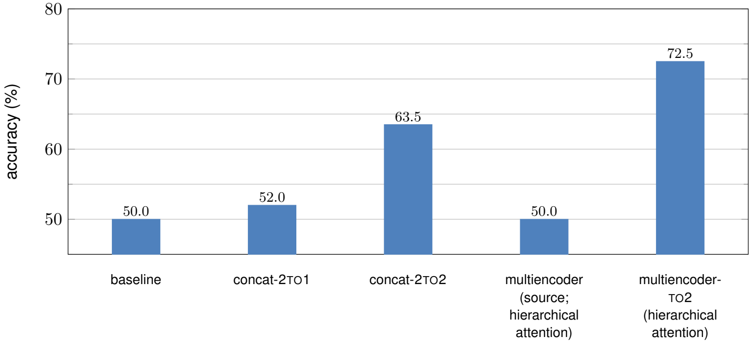
architecture exploration:

- condition on previous source, target, or both?
- use multiple encoders or just concatenate sentences?
- how to combine multiple context vectors in multi-encoder setups?
  - concatenate
  - gating mechanism
  - hierarchical attention

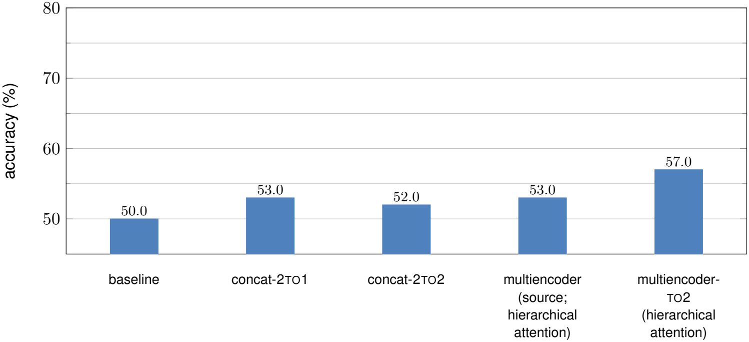
Results: BLEU



Results: Contrastive Test Set: Coreference



Results: Contrastive Test Set: Coherence/Cohesion



Large-Scale Evaluation: ContraPro

[Müller, Rios, Voita, Sennrich, WMT 2018]

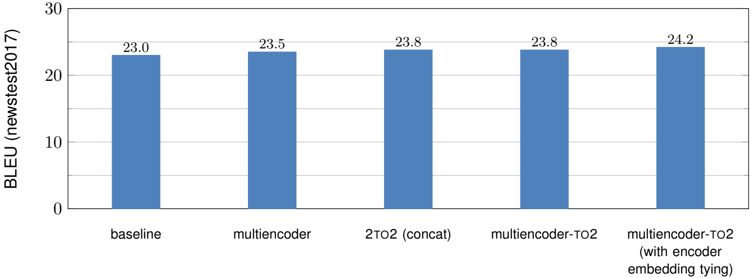


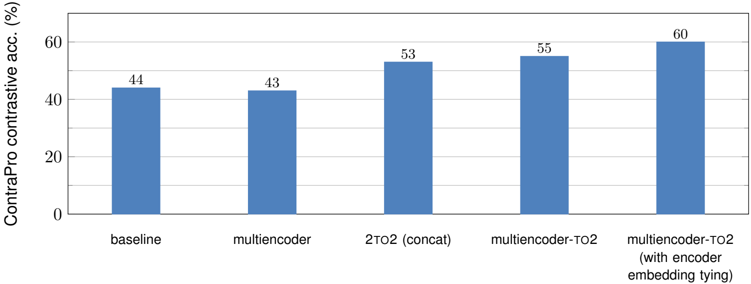

- 12 000 instances of ambiguous pronoun “it” (EN→DE)
  → German marks grammatical gender (3 classes) on all nouns
- real examples extracted from OpenSubtitles
- metadata for analysis of hard cases:
  - distant antecedents
  - minority classes

Research Questions

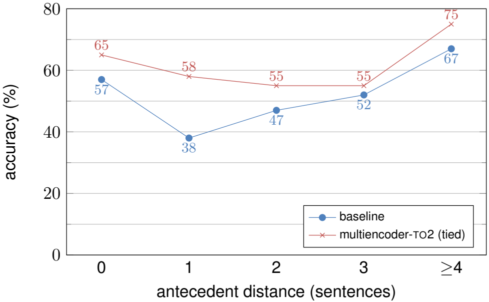
- can we confirm findings by [Bawden et al., 2018]
  on large-scale, more natural dataset?
- is training signal strong enough to learn good context encoder?
  Does parameter tying with main encoder help?

ContraPro: Selected Results





ContraPro: Interpreting Results

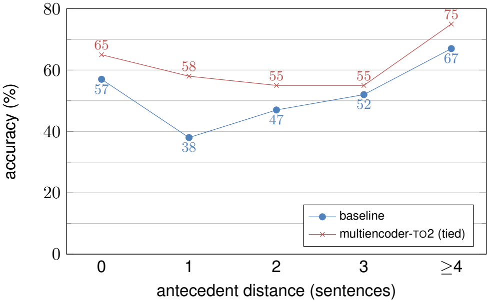


multiencoder-TO2 has context window of 1:

- why does quality improve when nominal antecedent is in same sentence, or further away?

- why does baseline improve with increased antecedent distance?

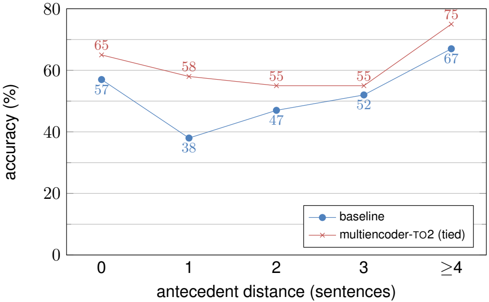
ContraPro: Interpreting Results



multiencoder-TO2 has context window of 1:

- why does quality improve when nominal antecedent is in same sentence, or further away?
  $\rightarrow$ coreference chains
- why does baseline improve with increased antecedent distance?

ContraPro: Interpreting Results



multiencoder-TO2 has context window of 1:

- why does quality improve when nominal antecedent is in same sentence, or further away?
  $\rightarrow$ coreference chains
- why does baseline improve with increased antecedent distance?
  $\rightarrow$ more instances of majority class

ContraPro: Coreference Chain Example

Example with antecedent distance 2:

	t-2	t-1	t
source (EN)	What's with **the door**?	**It** won't open.	- Is **it** locked?
target (DE)	Was ist mit **der Tür**?	**Sie** geht nicht auf.	- Ist **sie** abgeschlossen?

ContraPro: Conclusions

- confirms importance of target context for predicting agreement
- how context encoder is trained has big effect (weak learning signal?)
  - parameter tying between encoders helps [Voita et al., 2018]
  - promising direction: modify training objective [Jean and Cho, 2019]

When a Good Translation is Wrong in Context

[Voita, Sennrich, Titov, ACL 2019]

anaphora are well-known discourse phenomenon; what else do we find?

human evaluation:

- mark sentence-level translations as good or bad
- 2nd evaluation: if two consecutive translations are good, mark if they are also good in context of each other
- if translations are good in isolation, but not in context, annotate error
- data: English–Russian, OpenSubtitles

Human Evaluation of Consecutive Translations: Results

one/both bad	both good	
	bad pair	**good pair**
211	**140**	1649
11%	**7%**	82%

type of error	frequency
deixis	37%
ellipsis	29%
lexical cohesion	14%
ambiguity	9%
anaphora	6%
other	5%

Deixis

type of error	frequency
T-V distinction	67%
speaker/addressee gender:	
same speaker	22%
different speaker	9%
other	2%

translation errors caused by deixis (excluding anaphora)

Deixis

type of error	frequency
T-V distinction	67%
speaker/addressee gender:	
same speaker	22%
different speaker	9%
other	2%

translation errors caused by deixis (excluding anaphora)

EN We haven't really spoken much since your return. Tell me, what's on your mind these days?

RU Мы не разговаривали с тех пор, как вы вернулись. Скажи мне, что у тебя на уме в последнее время?

My ne razgovarivali s tekh por, kak vy vernulis'. Skazhi mne, chto u tebya na ume v posledneye vremya?

V-form (formal), T-form (informal)

Deixis

type of error	frequency
T-V distinction	67%
speaker/addressee gender:	
same speaker	22%
different speaker	9%
other	2%

translation errors caused by deixis (excluding anaphora)

EN I didn't come to Simon's for you. I did that for me.

RU Я пришла в Саймону не ради тебя. Я сделал это для себя.
Ya prishla v Saymonu ne radi tebya. Ya sdelal eto dlya sebya.

feminine, masculine.

Ellipsis

type of error	frequency
wrong morphological form	66%
wrong verb (VP-ellipsis)	20%
other error	14%

translation errors caused by ellipsis

Ellipsis

type of error	frequency
wrong morphological form	66%
wrong verb (VP-ellipsis)	20%
other error	14%

translation errors caused by ellipsis

EN You call her your friend but have you been to her home ? Her work ?

RU Ты называешь её своей подругой, но ты был у неё дома? Её работа?
Ty nazyvayesh’ yeyë svoyey podrugoy, no ty byl u neyë doma? Yeyë rabota?

wrong morphological form: noun phrase marked as subject

Ellipsis

type of error	frequency
wrong morphological form	66%
wrong verb (VP-ellipsis)	20%
other error	14%

translation errors caused by ellipsis

EN Veronica, thank you, but you saw what happened. We all did.

RU Вероника, спасибо, но ты видела, что произошло. Мы все хотели.
Veronika, spasibo, no ty videla, chto proizoshlo. My vse khoteli.

correct meaning is “see”, but MT produces хотели (“want”).

Lexical Cohesion

EN But that's not what <u>I'm talking about</u>. <u>I'm talking about</u> your future.

RU Но я говорю не об этом. Речь о твоём будущем.
No ya govoryu ne ob etom. Rech' o tvoyëm budushchem.

Inconsistent translation

EN Not for <u>Julia</u>. <u>Julia</u> has a taste for taunting her victims.

RU Не для Джулии. Юлия умеет дразнить своих жертв.
Ne dlya Dzhulii. Yuliya umeyet draznit' svoikh zhertv.

Name translation inconsistency

A Contrastive Test Set for Ellipsis, Deixis, and Lexical Cohesion

- held-out data from English–Russian OpenSubtitles
- *relevant context* up to 3 sentences away
- deixis: focus on T-V distinction
- lexical cohesion: focus on name translation consistency
- ellipsis:
  - predict NP inflection from context
  - predict verb from context

		latest relevant context		
	total	1st	2nd	3rd
deixis	3000	1000	1000	1000
lexical cohesion	2000	855	630	515
ellipsis (inflection)	500	500		
ellipsis (VP)	500	500		

Size of test sets

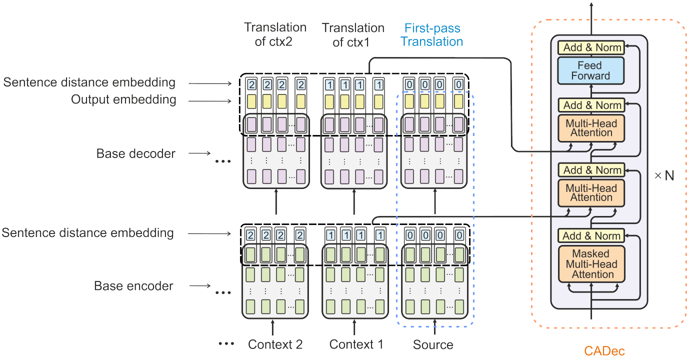
Research Questions

- how much does context-aware model help for deixis, ellipsis, lexical cohesion?
- how to build a context-aware model where most of the training data is sentence-level?

Training data

- OpenSubtitles English–Russian
- 6 million sentence pairs as starting point
- after data cleaning, 1.5 million sentence pairs have reliable context (1–3 sentences)

Model: Two-Pass Translation



Model architecture

Two-Pass Model

Training

- first-pass model is trained on all parallel data
- second-pass model is trained on subset with context
- second-pass model receives draft translation as input, either:
  - sampled from first-pass model
  - corrupted reference (20% of words randomly replaced)
- first-pass model is also used to compute hidden representations of current sentence and context

Inference

at test time, first-pass translation is produced with beam search

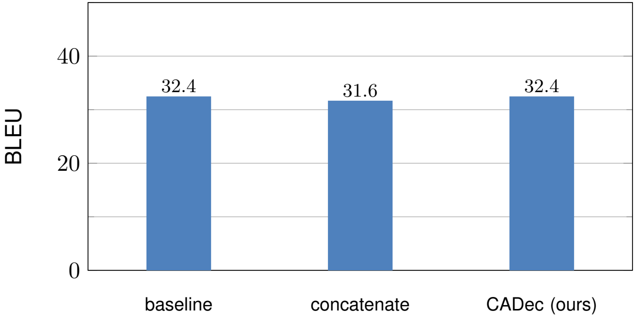
One-Pass Model

- concatenate sentences to form “context-aware” translation units
- train on mix of sentence-level and document-level data

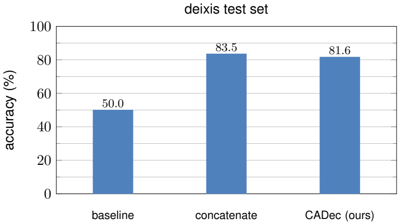
spoiler: this gave us poor BLEU results

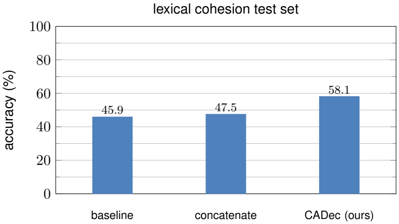
(other researchers had more success with pre-training model on sentence-level data, then fine-tuning on document-level data [Zhang et al., 2018, Tan et al., 2019])

Results: BLEU

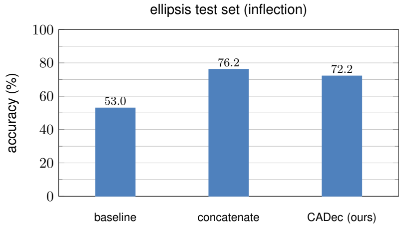


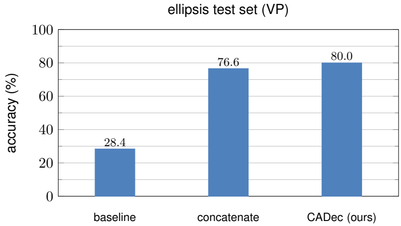
Contrastive Results





Contrastive Results





Results: Choice of First-Pass Translation During Training

p	**BLEU**	**deixis**	**lexical cohesion**	**ellipsis**
baseline	32.40	50.0	45.9	53 / 28
$p=0$	32.34	**84.1**	48.7	65 / 75
$p=0.25$	32.31	83.3	52.4	67 / 78
$p=0.5$	32.38	81.6	58.1	**72** / **80**
$p=0.75$	32.45	80.0	**65.0**	70 / **80**

Results for different probabilities of using corrupted reference at training time.
BLEU for 3 context sentences. For ellipsis, we show inflection/VP scores.

Changes with small effect on BLEU can have large effect on consistency!

Using Monolingual Document-level Data

document context is often lost in parallel data extraction

what can we do if **all** parallel data is sentence-level, and we only have monolingual data with wider context?

solution 1: noisy channel model [Yu et al., 2019]

$T^* = \arg\max_T P(S|T)P(T)$

- channel model ($P(S|T)$) operates on sentence-level.
- language model ($P(T)$) operates on document-level.

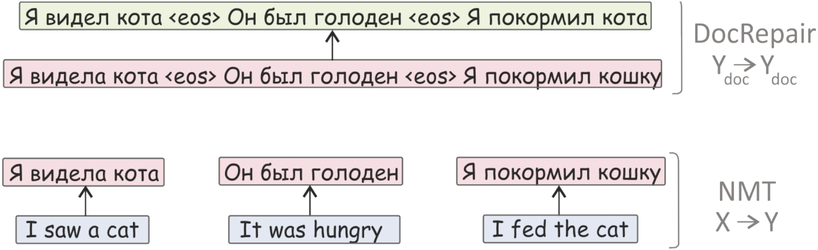
solution 2: automatic post-editing (monolingual repair)

Context-Aware Monolingual Repair

[Voita, Sennrich, Titov, EMNLP 2019]



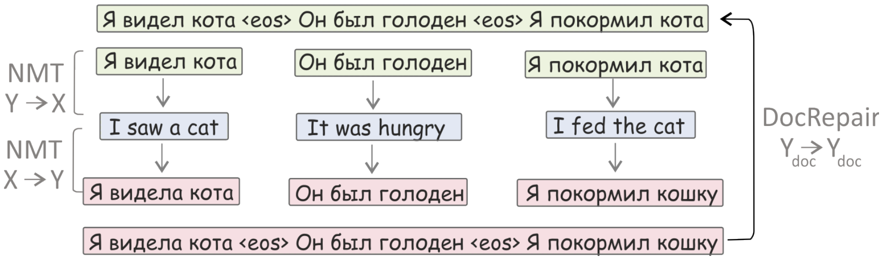
1. translate sentences independently
2. fix inconsistencies with multi-sentence monolingual repair model



Training Monolingual Repair Model

how to train monolingual repair model?

- simple sequence-to-sequence model with Transformer
- target side: original text in target language
- source side: original text, translated to source language and back with sentence-level system



Some Results [Voita et al., 2019b, Voita et al., 2019a]

system	BLEU	consistency test sets			
		deixis	lexical cohesion	ellipsis (infl.)	ellipsis (VP)
sentence-level	33.9	50.0	45.9	53.0	28.4
concatenation (4-to-4)	-	83.5	47.5	76.2	76.6
CADec	-	81.6	58.1	72.2	**80.0**
monolingual repair	**34.6**	**91.8**	**80.6**	**86.4**	75.2

Discussion

- monolingual repair best in terms of BLEU, and most contrastive test sets
- why poorer performance for VP ellipsis?
  → fewer VP ellipses in synthetic source sentences

(a)	**EN**	No one believed me. But she did.
	RU	Мне никто не верил. Но она сказала.
(b)	**RU**	Никто мне не верил. Но она верила.
	EN	No one believed me. But she believed.
	RU	Мне никто не верил. Но она поверила.

real source sentence (a) vs. synthetic example (b)

Conclusions

- sentence-level machine translation is not “good enough”
- context-aware models have large effects...
  ...but we need tools to better measure them
- targeted evaluation shows effect of context-aware models:
  $\rightarrow$ small design decisions have big impact on ”context-awareness“!
- monolingual models are attractive because of data requirements and potential applications

Thank you for your attention

Resources

- English–French contrastive test set:
  https://diamt.limsi.fr/eval.html
- large-scale contrastive test set of context-aware pronoun translation:
  https://github.com/ZurichNLP/ContraPro
- code and data for English–Russian experiments:
  https://github.com/lena-voita/good-translation-wrong-in-context

Bibliography I

Bahdanau, D., Cho, K., and Bengio, Y. (2015).
Neural Machine Translation by Jointly Learning to Align and Translate.
In Proceedings of the International Conference on Learning Representations (ICLR).

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018).
Evaluating Discourse Phenomena in Neural Machine Translation.
In NAACL 2018, New Orleans, USA.

Guillou, L. and Hardmeier, C. (2018).
Automatic reference-based evaluation of pronoun translation misses the point.
In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4797–4802, Brussels, Belgium. Association for Computational Linguistics.

Jean, S. and Cho, K. (2019).
Context-Aware Learning for Neural Machine Translation.
CoRR, abs/1903.04715.

Jean, S., Lauly, S., Firat, O., and Cho, K. (2017).
Neural Machine Translation for Cross-Lingual Pronoun Prediction.
In Proceedings of the 3rd Workshop on Discourse in Machine Translation, DISCOMT'17, pages 54–57, Copenhagen, Denmark.

Junczys-Dowmunt, M. (2019).
Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation.
In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 225–233, Florence, Italy. Association for Computational Linguistics.

Müller, M., Rios, A., Voita, E., and Sennrich, R. (2018).
A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation.
In Proceedings of the Third Conference on Machine Translation, pages 61–72, Belgium, Brussels.

Bibliography II

Tan, X., Zhang, L., Xiong, D., and Zhou, G. (2019).
Hierarchical modeling of global context for document-level neural machine translation.
In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1576–1585, Hong Kong, China. Association for Computational Linguistics.

Tiedemann, J. and Scherrer, Y. (2017).
Neural Machine Translation with Extended Context.
In Proceedings of the Third Workshop on Discourse in Machine Translation, pages 82–92, Copenhagen, Denmark.

Voita, E., Sennrich, R., and Titov, I. (2019a).
Context-aware monolingual repair for neural machine translation.
In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 876–885, Hong Kong, China. Association for Computational Linguistics.

Voita, E., Sennrich, R., and Titov, I. (2019b).
When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion.
In Proceedings of the 57th Conference of the Association for Computational Linguistics, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Voita, E., Serdyukov, P., Sennrich, R., and Titov, I. (2018).
Context-Aware Neural Machine Translation Learns Anaphora Resolution.
In ACL 2018, Melbourne, Australia.

Bibliography III

Wang, L., Tu, Z., Way, A., and Qun Liu (2017).
Exploiting Cross-Sentence Context for Neural Machine Translation.
In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP'17, pages 2816–2821, Denmark, Copenhagen.

Wong, B. T. M. and Kit, C. (2012).
Extending machine translation evaluation metrics with lexical cohesion to document level.
In
Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Langua
pages 1060–1068, Jeju Island, Korea. Association for Computational Linguistics.

Yu, L., Sartran, L., Stokowiec, W., Ling, W., Kong, L., Blunsom, P., and Dyer, C. (2019).
Putting machine translation in context with the noisy channel model.

Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M., and Liu, Y. (2018).
Improving the transformer translation model with document-level context.
In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.