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

Rico Sennrich

University of Edinburgh

Achieving Human Parity

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

March 14, 2018 | Allison Linn

Achieving Human Parity

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

March 14, 2018 | Allison Linn

laudable...

- follows best practices with WMT-style evaluation
- data released for scientific scrutiny (outputs, references, rankings)

Achieving Human Parity

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

March 14, 2018 | Allison Linn

...but warrants further scrutiny

- failure to reject null hypothesis is not evidence of parity
- alternative hypothesis:
  human raters prefer human translations on a **document-level**
- rationale:
  - context helps raters understand text and spot semantic errors
  - discourse errors are invisible in sentence-level evaluation

A Case for Document-level Evaluation

[Läubli, Sennrich, Volk, EMNLP 2018]



can we reproduce Microsoft's finding with different evaluation protocol?

	original evaluation	our evaluation
test set	WMT17	WMT17 (native Chinese part)
system	Microsoft COMBO-6	Microsoft COMBO-6
raters	crowd-workers	**professional translators**
experimental unit	sentence	sentence / **document**
measurement	direct assessment	**pairwise ranking**
raters see reference	no	no
raters see source	yes	yes / no
ratings	$\geq$ 2,520 per system	$\approx$ 200 per setting

Which Text is Better?

Members of the public who find their cars obstructed by unfamiliar vehicles during their daily journeys can use the "Twitter Move Car" feature to address this distress when the driver of the unfamiliar vehicle cannot be reached.

A citizen whose car is obstructed by vehicle and is unable to contact the owner of the obstructing vehicle can use the "WeChat Move the Car" function to address the issue.

Which Text is Better?

Members of the public who find their cars obstructed by unfamiliar vehicles during their daily journeys can use the "Twitter Move Car" feature to address this distress when the driver of the unfamiliar vehicle cannot be reached.

On August 11, Xi'an traffic police WeChat service number "Xi'an traffic police" launched "WeChat mobile" service.

With the launch of the service, members of the public can tackle such problems in their daily lives by using the "WeChat Move" feature when an unfamiliar vehicle obstructs the movement of their vehicle while the driver is not at the scene. [...]

A citizen whose car is obstructed by vehicle and is unable to contact the owner of the obstructing vehicle can use the "WeChat Move the Car" function to address the issue.

The Xi'an Traffic Police WeChat official account "Xi'an Jiaojing" released the "WeChat Move the Car" service since August 11.

Once the service was released, a fellow citizen whose car was obstructed by another vehicle and where the driver of the vehicle was not present, the citizen could use the "WeChat Move the Car" function to address the issue. [...]

Which Text is Better?

市民在日常出行中,发现爱车被陌生车辆阻碍了,在联系不上陌生车辆司机的情况下,可以使用“微信挪车”功能解决这一困扰。

8月11日起,西安交警微信服务号“西安交警”推出“微信挪车”服务。

这项服务推出后,日常生活中,市民如遇陌生车辆在驾驶人不在现场的情况下阻碍自己车辆行驶时,就可通过使用“微信挪车”功能解决此类问题。[...]

Members of the public who find their cars obstructed by unfamiliar vehicles during their daily journeys can use the "Twitter Move Car" feature to address this distress when the driver of the unfamiliar vehicle cannot be reached.

On August 11, Xi'an traffic police WeChat service number "Xi'an traffic police" launched "WeChat mobile" service.

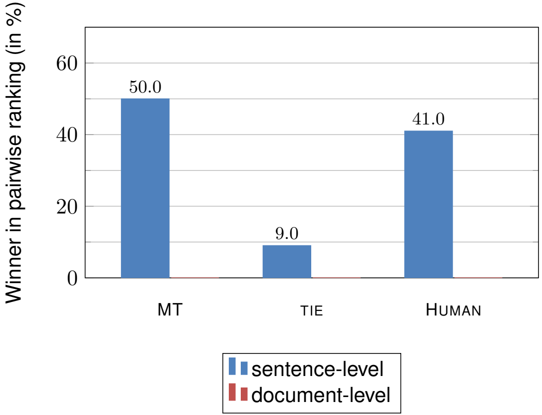
With the launch of the service, members of the public can tackle such problems in their daily lives by using the "WeChat Move" feature when an unfamiliar vehicle obstructs the movement of their vehicle while the driver is not at the scene. [...]

A citizen whose car is obstructed by vehicle and is unable to contact the owner of the obstructing vehicle can use the "WeChat Move the Car" function to address the issue.

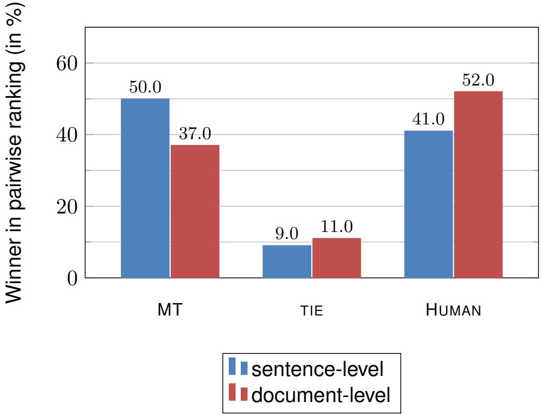
The Xi'an Traffic Police WeChat official account "Xi'an Jiaojing" released the "WeChat Move the Car" service since August 11.

Once the service was released, a fellow citizen whose car was obstructed by another vehicle and where the driver of the vehicle was not present, the citizen could use the "WeChat Move the Car" function to address the issue. [...]

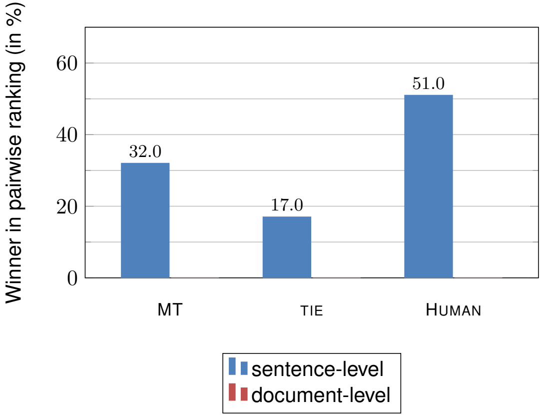
Evaluation Results: Adequacy Assessment



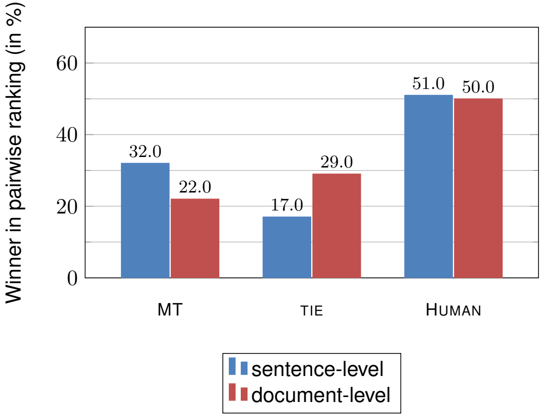
Evaluation Results: Adequacy Assessment



Evaluation Results: Fluency Assessment

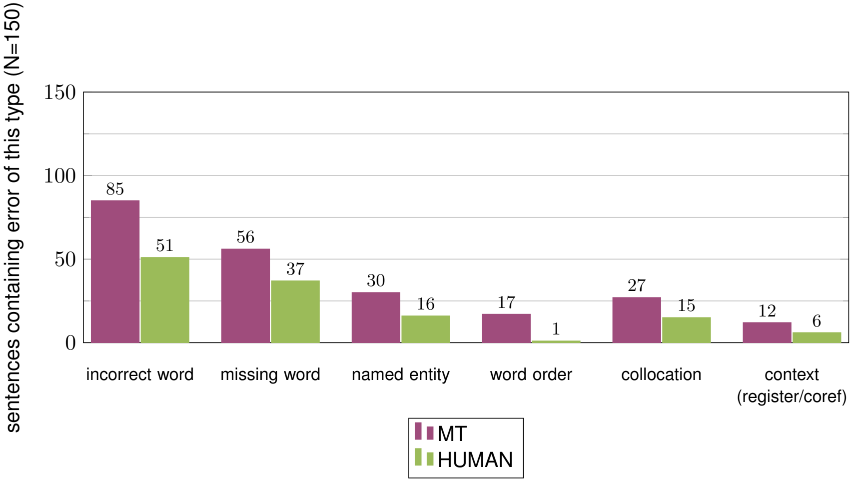


Evaluation Results: Fluency Assessment



Follow-Up Study: Error Analysis

[Läubli, Castilho, Neubig, Sennrich, Shen, Toral, in preparation]



Human Evaluation Results

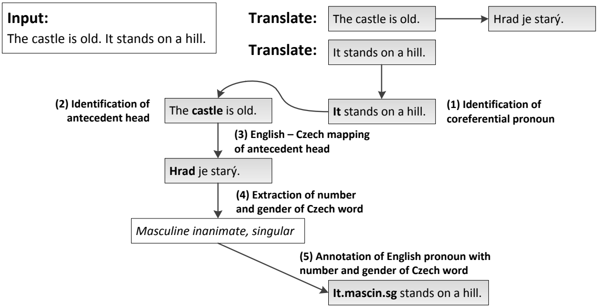
- document-level ratings show significant preference for HUMAN
- preference for HUMAN is even stronger in fluency evaluation
- error analysis shows MT makes more errors, partially related to context and consistency

Conclusions

- discourse-level cohesion and coherence is important, but invisible in sentence-level evaluation
- distinguishing MT from human translations becomes harder with increasing quality
  $\rightarrow$ WMT 2019 is shifting to document-level evaluation

New Chances: Context-Aware NMT

SMT era:

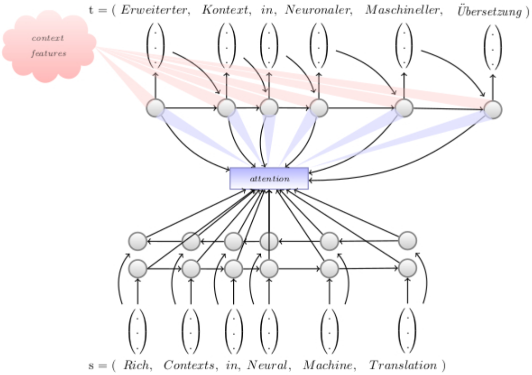


specialized features

[Hardmeier, 2012, Guillou, 2012, Meyer et al., 2012]

New Chances: Context-Aware NMT

NMT era:



contextual sentences as additional input

[Jean et al., 2017, Wang et al., 2017, Tiedemann and Scherrer, 2017, Bawden et al., 2018, Voita et al., 2018, Maruf and Haffari, 2018]

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?

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

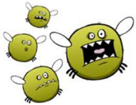
A Contrastive Test Set: Coreference

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!

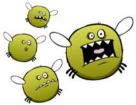
sentence: Don’t worry, I'll kill **it** for you.

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?

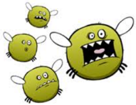
Previous linguistic context necessary to disambiguate

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

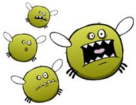
Can the model rank the correct sentence above the incorrect one?

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:

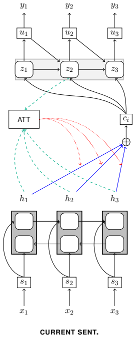
context:	What's **crazy** about me?
current sent.:	Is this **crazy**?

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 ?

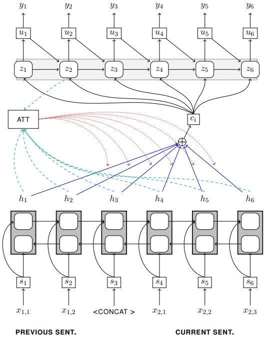
Architectures

Baseline



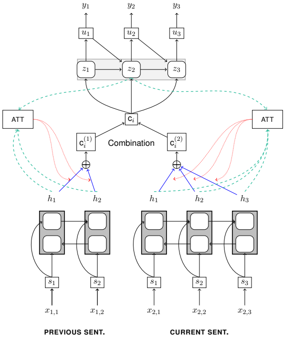
[Bahdanau et al., 2015]

2TO2 - concatenated input



[Tiedemann and Scherrer, 2017]

Multiple encoders



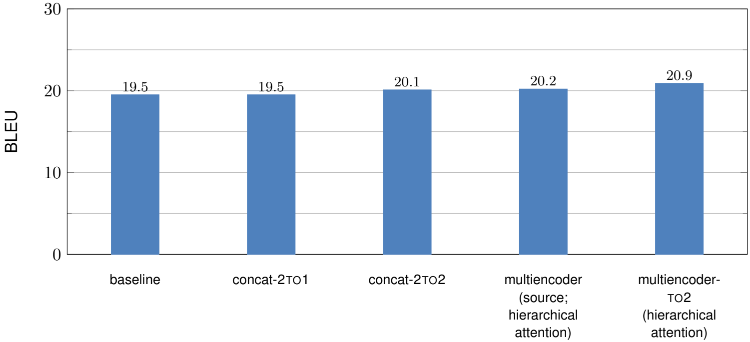
[Jean et al., 2017, Wang et al., 2017]

Architectures

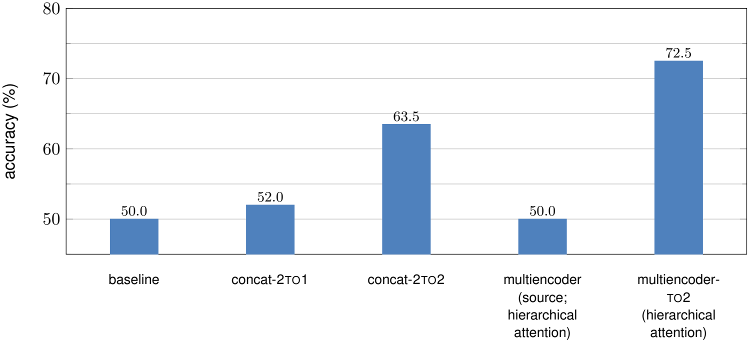
architecture exploration:

- 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

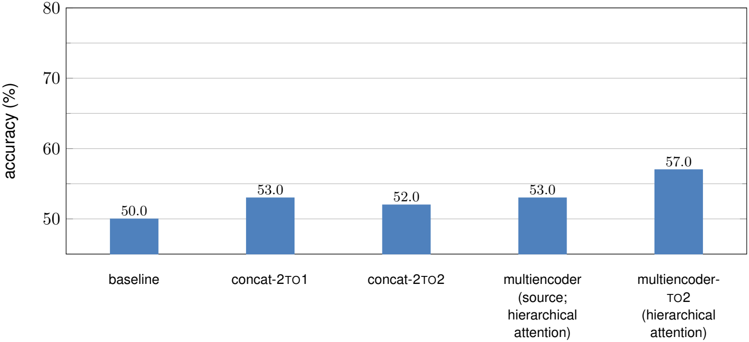
Results: BLEU



Results: Contrastive Test Set: Coreference



Results: Contrastive Test Set: Coherence/Cohesion



Large-Scale Evaluation: ContraPro

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

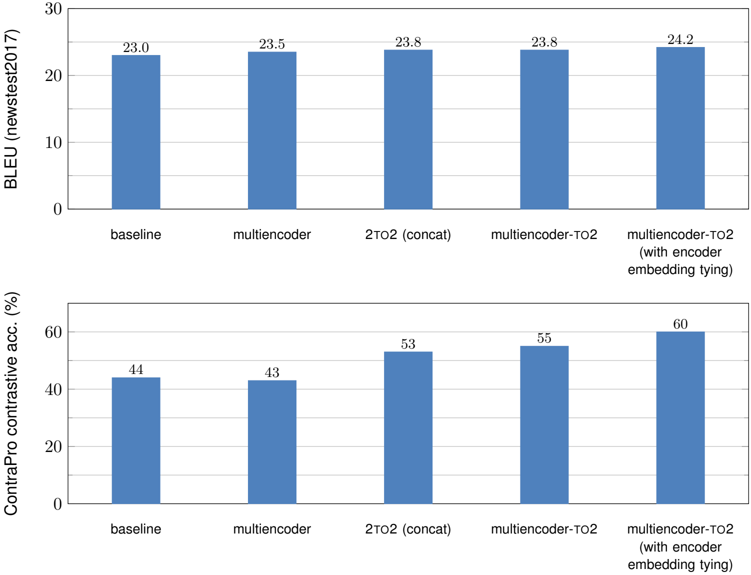
   

- 12 000 instances of ambiguous pronoun “it” (EN→DE)
  → German marks grammatical gender (3 classes) on all nouns
- real examples extracted from OpenSubtitles
- metadata for analysis of hard cases:
  - distant antecedents
  - minority classes

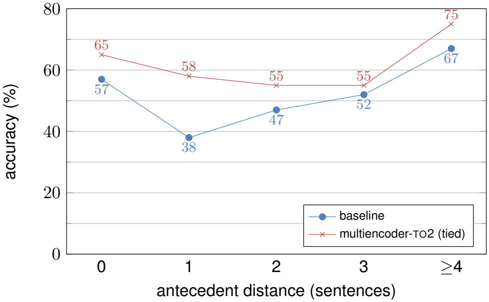
Research Questions

- 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

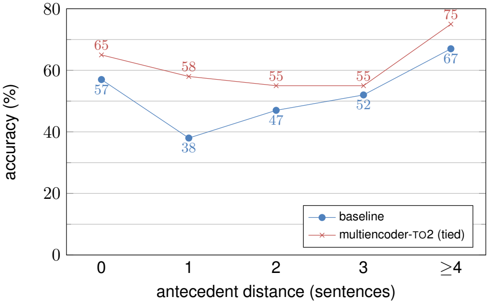


multiencoder-TO2 has context window of 1:

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

- why does baseline improve with increased antecedent distance?

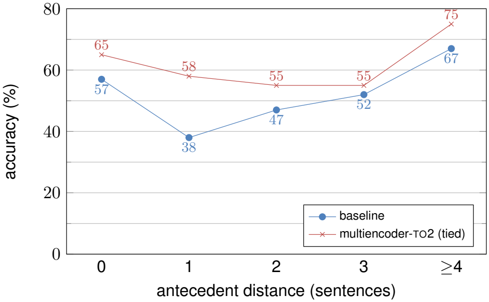
ContraPro: Interpreting Results



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, in preparation]

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 I'm talking about. I'm talking about 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 Julia. Julia 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

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?

BERT-produced text

human-produced text

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

can we distinguish accidental repetition from document-level cohesion?

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

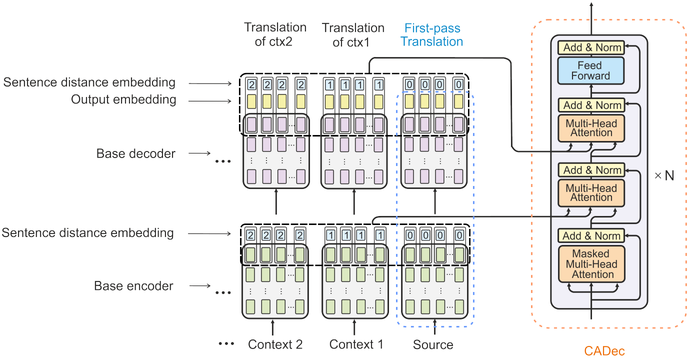
Research Questions

- 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

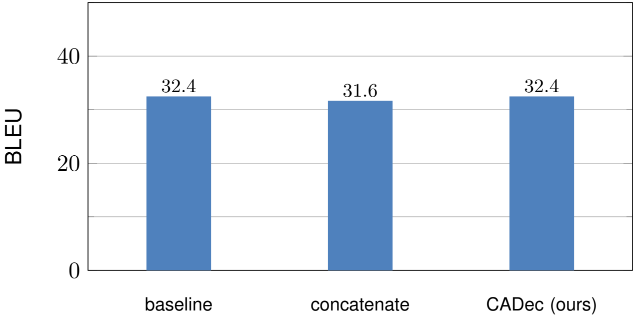
Training

- 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

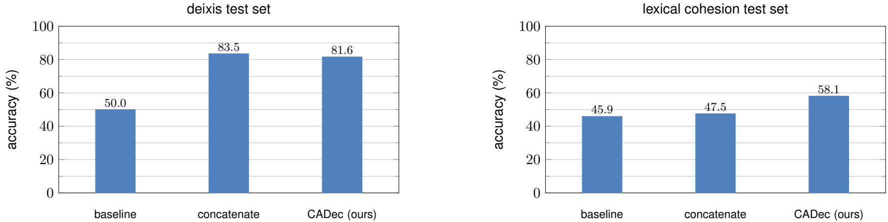
Inference

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

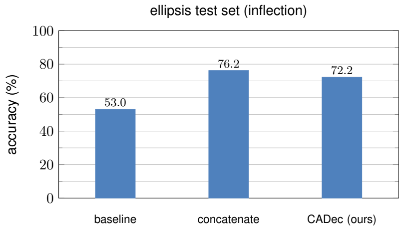
Results: BLEU

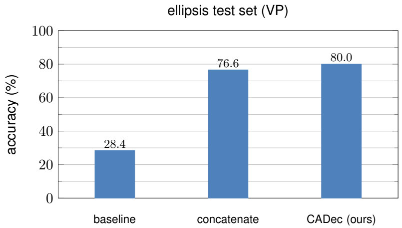


Contrastive Results



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!

Open Question: Which Context Matters?

most work so far focuses on previous sentence(s), but:

- relevant information can be further in past
- relevant information can be in future context

source	I went there with **my friend**. **She** was amazed to see that it had multiple floors.
reference	Sono andato la' con **la mia amica**. E' rimasta meraviglia nel vedere che aveva piu' piani
baseline	Arrivai li con **il mio amico**. Rimaneva meravigliato di vedere che aveva una cosa piu incredibile.
contextual (prev+next)	Sono andato con **la mia amica**. Fu sorpresa nel vedere che aveva piu piani.

[Agrawal et al., 2018]

Outlook: WMT 2019

effort to move training data and human evaluation to document level

System Output List

English-German newstest2019

Translations | Resources | Download | Info | Account

Interested in Contributing?

- Check out available resources.
- Create an account and start submitting your own systems.

Scored Systems

System	Submitter	System Notes	Constraint	Run Notes	BLEU	BLEU-cased
Microsoft-WMT19-sentence/document *(Details)*	marcinjd *Microsoft*	Sentence-level/document-level combination via 2-pass decoding from sentence to document-level.	yes		49.0	**48.6**
MSRA.MADL *(Details)*	MSRA-WXW *Microsoft*		yes		48.8	48.5
Microsoft-WMT19-document-level *(Details)*	marcinjd *Microsoft*	Pure document-level system	yes		48.7	48.4
Facebook FAIR *(Details)*	edunov *Facebook FAIR*		yes	+ fix-quotes	47.9	47.5
NEU *(Details)*	NiuTrans *Northeastern University*	Ensemble of 8 deep Transformer (30 layers) models + back-translation with sampling + distillation by ensemble teachers + hypothesis combination + fix quotes	yes		47.3	47.0
UCAM *(Details)*	fstahlberg *University of Cambridge*		yes	1 sentence-level LM, 1 document-level LM, 4 NMT models fine-tuned with EWC, fixed quotes	46.5	46.1
Microsoft-WMT19-sentence-level *(Details)*	marcinjd *Microsoft*	Pure sentence-level system	yes		46.3	46.0
JHU *(Details)*	kelly-yash-jhu *Johns Hopkins University*		yes	Fix quotes	45.5	45.1
Microsoft WMT18-baseline *(Details)*	marcinjd *Microsoft*		yes	WMT18-baseline, same as last year, fixed quotes.	45.2	44.8
Helsinki-NLP *(Details)*	aarnetalman *University of Helsinki*		yes	Transformer ensemble	45.1	44.5
eTranslation *(Details)*	eTranslation *DGT/CNECT*	base transformer ensemble of 3 models plus LM, fine tuned on devset	yes	FQ	44.8	44.4
lmu-ctx-tf-single-en-de *(Details)*	dario *LMU Munich*	context-aware single transformer big model	yes	fixed quotes	43.8	43.4
MLLP-UPV *(Details)*	mllp *MLLP group - Univ. Politècnica de València*	Transformer big model. Includes 10M sentences from Paracrawl and 18M backtranslated sentences. Finetuned on newstest08-16. Single model.	yes		43.7	43.3
dfki-nmt *(Details)*	zhangjingyi *dfki*		yes		44.3	43.0
dfki-nmt *(Details)*	zhangjingyi *dfki*		yes		43.3	42.0
test *(Details)*	kxyg *SYU*	test	yes		42.2	41.8

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“!

Thank you for your attention

Resources

- Evaluation data on human parity:
  `https://github.com/laeubli/parity`
- contrastive test sets for discourse in MT evaluation:
  `https://github.com/rbawden/discourse-mt-test-sets`
- large-scale contrastive test set of context-aware pronoun translation:
  `https://github.com/ZurichNLP/ContraPro`

Bibliography I

Agrawal, R., Turchi, M., and Negri, M. (2018).
Contextual Handling in Neural Machine Translation: Look Behind, Ahead and on Both Sides.
In 21th Annual Conference of the European Association for Machine Translation.

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. (2012).
Improving Pronoun Translation for Statistical Machine Translation.
In
Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computation
pages 1–10, Avignon, France.

Hardmeier, C. (2012).
Discourse in statistical machine translation: A survey and a case study.
Discours, 11.

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.

Bibliography II

Läubli, S., Sennrich, R., and Volk, M. (2018).
Has Neural Machine Translation Achieved Human Parity? A Case for Document-level Evaluation.
In EMNLP 2018, Brussels, Belgium.

Maruf, S. and Haffari, G. (2018).
Document Context Neural Machine Translation with Memory Networks.
In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1275–1284.

Meyer, T., Popescu-Belis, A., Hajlaoui, N., and Gesmundo, A. (2012).
Machine Translation of Labeled Discourse Connectives.
In Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA).

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.

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

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.

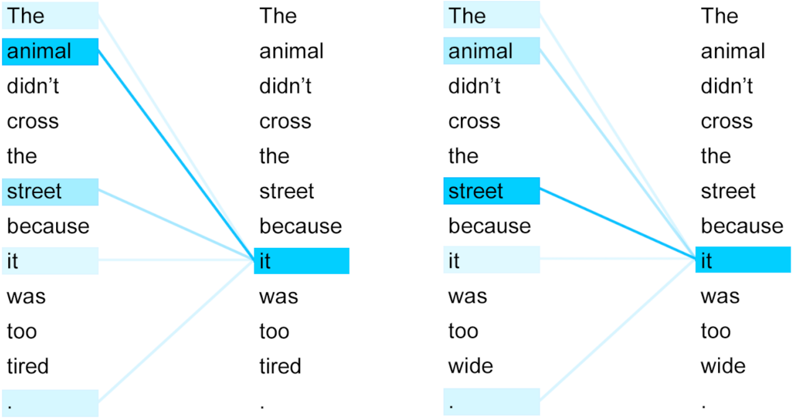
Bibliography III

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.

Analyzing Use of Context: RNN

Figure 6: Attention patterns with referential pronouns in extended context.

Analyzing Use of Context: Transformer



[?]

Analyzing Use of Context

[Voita, Serdyukov, Sennrich, Titov, ACL 2018]




set-up

- Transformer architecture with clear interface to context
- analysis of attention patterns

Context-Aware Transformer

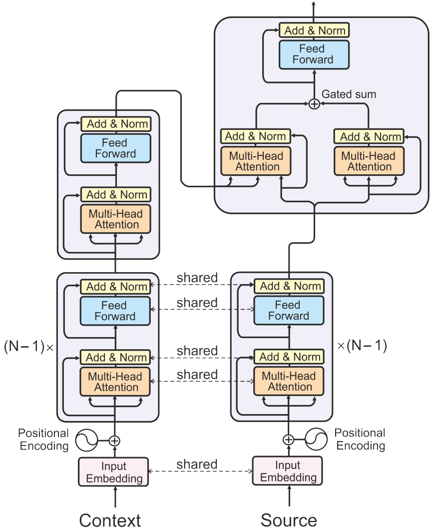
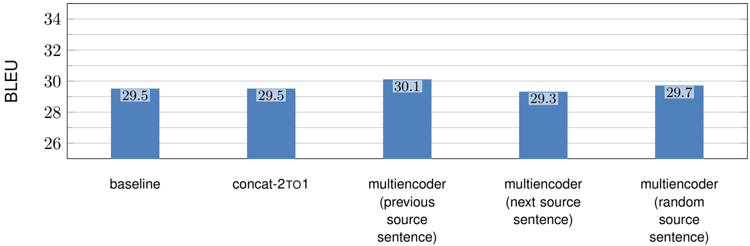


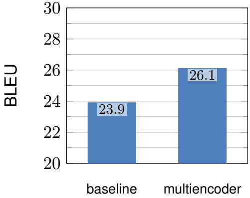
Figure 1: Encoder of the discourse-aware model

Context-Aware Transformer: Evaluation

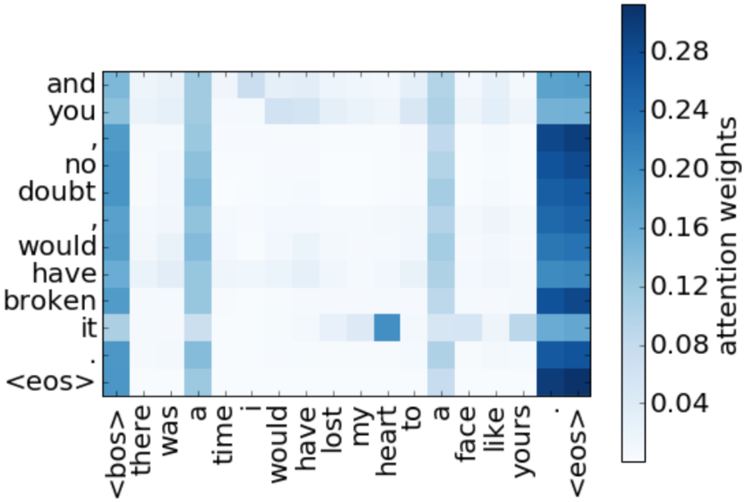
- OpenSubtitles2018 English→Russian
- scores on random test set:



larger improvements on focused test set ('it' with nominal antecedent):



Context-Aware Transformer Learns Anaphora Resolution



Context-Aware Transformer Learns Anaphora Resolution

	agreement (in %)
coreNLP	77
attention	72
last noun	54

Agreement with human assessment for coreference resolution of anaphoric *it*.
Examples with ≥ 1 noun in context sentence.

Examples from Top WMT18 Systems

coreference

In fairness, Miller did not attack the statue itself.
[...]
But he did attack its meaning [...]

HUMAN	MT
Um fair zu bleiben, Miller griff nicht die Statue selbst an. [...] Aber er griff deren Bedeutung an [...]	Fairerweise hat Miller die Statue nicht selbst angegriffen. [...] Aber er griff seine Bedeutung an [...]

Examples from Top WMT18 Systems

coreference

In fairness, Miller did not attack the statue itself.
[...]
But he did attack its meaning [...]

HUMAN	MT
Um fair zu bleiben, Miller griff nicht die Statue selbst an. [...] Aber er griff deren Bedeutung an [...]	Fairerweise hat Miller die Statue nicht selbst angegriffen. [...] Aber er griff seine Bedeutung an [...]

Examples from Top WMT18 Systems

lexical coherence

Weidezaunprojekt ist elementar

Das Fischerbacher Weidezaun-Projekt ist ein Erfolgsprojekt und wird im kommenden Jahr fortgesetzt.

HUMAN	MT
Pasture fence project is fundamental	Electric fence project is basic
The Fischerbach pasture fence project is a successful project and will be continued next year.	The Fischerbacher Weidezaun-Project is a success and will be continued in the coming year.

Examples from Top WMT18 Systems

lexical coherence

Weidezaunprojekt ist elementar

Das Fischerbacher Weidezaun-Projekt ist ein Erfolgsprojekt und wird im kommenden Jahr fortgesetzt.

HUMAN	MT
Pasture fence project is fundamental	Electric fence project is basic
The Fischerbach pasture fence project is a successful project and will be continued next year.	The Fischerbacher Weidezaun-Project is a success and will be continued in the coming year.

Examples from Top WMT18 Systems

pro-drop

该款机器人使用语音合成、[...]

曾获得国际消费电子产品展（CES）[...]

HUMAN	MT
This robot uses speech synthesis, [...] with conversational [...] features.	Using speech synthesis [...] the robot has the functions of chatting conversation [...]
It has won two major CES awards [...]	Has won two awards at the International Consumer Electronics Exhibition (CES) [...]

Examples from Top WMT18 Systems

pro-drop

该款机器人使用语音合成、[...]

曾获得国际消费电子产品展（CES）[...]

HUMAN	MT
This robot uses speech synthesis, [...] with conversational [...] features.	Using speech synthesis [...] the robot has the functions of chatting conversation [...]
It has won two major CES awards [...]	Has won two awards at the International Consumer Electronics Exhibition (CES) [...]