How Contextual is Neural Machine Translation?

Rico Sennrich

University of Zurich[UZH]

Edinburgh NLP
University of Edinburgh
Natural Language Processing

Machine Translation and the Limits of Generic Knowledge

Er	hat	einen	Krebstest	entwickelt

a caricature of rule-based MT:

Machine Translation and the Limits of Generic Knowledge

Er	hat	einen	Krebstest	entwickelt
he	?	?	?	?

a caricature of rule-based MT:

- let's translate a sentence word-by-word via a bilingual dictionary ☺

Machine Translation and the Limits of Generic Knowledge

Er	hat	einen	Krebstest	entwickelt
he	have	a	?	develop

a caricature of rule-based MT:

- let's translate a sentence word-by-word via a bilingual dictionary ☺
- hm, we need a morphology tool to deal with inflected forms... ☺

Machine Translation and the Limits of Generic Knowledge

Er	hat	einen	Krebstest	entwickelt
he	have	a	crab test	develop

a caricature of rule-based MT:

- let's translate a sentence word-by-word via a bilingual dictionary ☺
- hm, we need a morphology tool to deal with inflected forms... ☺
- ...and with compounds and derivational morphology ☺

Machine Translation and the Limits of Generic Knowledge

Er	hat	einen	Krebstest	entwickelt
he	has	a	crab test	develops

a caricature of rule-based MT:

- let's translate a sentence word-by-word via a bilingual dictionary ☺
- hm, we need a morphology tool to deal with inflected forms... ☺
- ...and with compounds and derivational morphology ☺
- oh, and we need to transfer and generate morphological features 😐

Machine Translation and the Limits of Generic Knowledge

Er hat einen Krebstest entwickelt

he has developed a crab test

a caricature of rule-based MT:

- let's translate a sentence word-by-word via a bilingual dictionary ☺
- hm, we need a morphology tool to deal with inflected forms... ☺
- ...and with compounds and derivational morphology ☺
- oh, and we need to transfer and generate morphological features 😐
- actually, we need syntactic transfer for disambiguation and restructuring ☹

Machine Translation and the Limits of Generic Knowledge

Er	hat	einen	Krebstest	entwickelt
he	has	developed	a	crab test

a caricature of rule-based MT:

- let's translate a sentence word-by-word via a bilingual dictionary ☺
- hm, we need a morphology tool to deal with inflected forms... ☺
- ...and with compounds and derivational morphology ☺
- oh, and we need to transfer and generate morphological features 😐
- actually, we need syntactic transfer for disambiguation and restructuring ☹
- wait, how are we going to disambiguate "Krebs" with rules? ☹

Krebs *m* (*genitive* **Krebses**, *plural* **Krebse**)

1. crab
2. cancer (disease)
3. (*astronomy*, *astrology*) Cancer

Today's Talk

How Contextual is Neural Machine Translation?

- a success story:
  word sense disambiguation based on sentence context
- an open challenge:
  co-reference across sentences

Word Sense Disambiguation

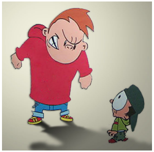
system	sentence
source	Dort wurde er von dem **Schläger** und einer weiteren männl. Person erneut angegriffen.
reference	There he was attacked again by his **original attacker** and another male.
our NMT	There he was attacked again by the **racket** and another male person.
Google	There he was again attacked by the **bat** and another male person.

Schläger

Word Sense Disambiguation

system	sentence
source	Dort wurde er von dem **Schläger** und einer weiteren männl. Person erneut angegriffen.
reference	There he was attacked again by his **original attacker** and another male.
our NMT	There he was attacked again by the **racket** and another male person.
Google	There he was again attacked by the **bat** and another male person.

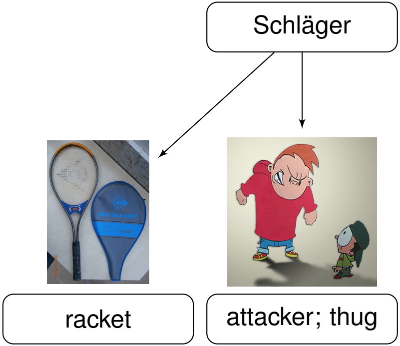
Schläger



attacker; thug

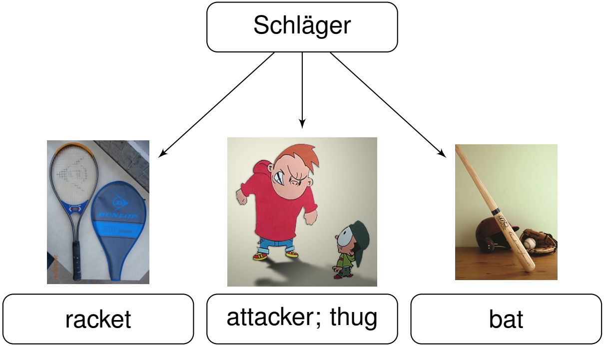
Word Sense Disambiguation

system	sentence
source	Dort wurde er von dem **Schläger** und einer weiteren männl. Person erneut angegriffen.
reference	There he was attacked again by his **original attacker** and another male.
our NMT	There he was attacked again by the **racket** and another male person.
Google	There he was again attacked by the **bat** and another male person.



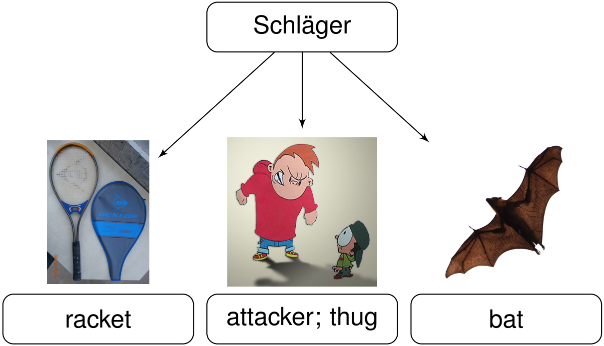
Word Sense Disambiguation

system	sentence
source	Dort wurde er von dem **Schläger** und einer weiteren männl. Person erneut angegriffen.
reference	There he was attacked again by his **original attacker** and another male.
our NMT	There he was attacked again by the **racket** and another male person.
Google	There he was again attacked by the **bat** and another male person.



Word Sense Disambiguation

system	sentence
source	Dort wurde er von dem **Schläger** und einer weiteren männl. Person erneut angegriffen.
reference	There he was attacked again by his **original attacker** and another male.
our NMT	There he was attacked again by the **racket** and another male person.
Google	There he was again attacked by the **bat** and another male person.



Neural Machine Translation in 2016: So Far, So Bad

source	*We thought a win like this might be close$_{adj}$.*
reference	*Wir dachten, dass ein solcher Sieg nah sein könnte.*
NMT (uedin WMT16)	**Wir dachten, ein Sieg wie dieser könnte schließen.*

Generic Knowledge and Word Sense Disambiguation

Etymology 1

From Old English *clȳsan* (“to close, shut”)

Verb

close (*third-person singular simple present* **closes**, *present participle* **closing**, *simple past and past participle* **closed**)

1. (*physical*) To remove a gap. DE: schließen
2. (*social*) To finish, to terminate. DE: beenden

Noun

close (*plural* **closes**)

1. An end or conclusion. DE: Ende

Etymology 2

Borrowed from French *clos*, from Latin *clausum*, participle of *claudō*.

Adjective

close (*comparative* **closer**, *superlative* **closest**)

1. Narrow; confined. DE: eng
2. At a little distance; near. DE: nah

Noun

close (*plural* **closes**)

1. (*chiefly British*) A street that ends in a dead end. DE: Sackgasse
2. (*Scotland*) A very narrow alley between two buildings, often overhung by one of the buildings above the ground floor. DE: Gasse

Adding Linguistic Knowledge to Neural MT
[Sennrich, Haddow, WMT 2016]

syntactic information in embedding

$$E_1(close) = \begin{bmatrix} 0.4 \\ 0.1 \\ 0.2 \end{bmatrix} \qquad E_2(adj) = \begin{bmatrix} 0.1 \end{bmatrix}$$

$$E_1(close) \parallel E_2(adj) = \begin{bmatrix} 0.4 \\ 0.1 \\ 0.2 \\ 0.1 \end{bmatrix}$$

source	*We thought a win like this might be close$_{adj}$.*
reference	*Wir dachten, dass ein solcher Sieg nah sein könnte.*
NMT (uedin WMT16)	*Wir dachten, ein Sieg wie dieser könnte schließen.*
+POS, dependency, lemma, morphology	*Wir dachten, ein Sieg wie dieser könnte nah sein.*

Evaluating WSD in MT
[Rios, Mascarell, Sennrich, WMT 2017]
[Rios, Müller, Sennrich, WMT 2018]

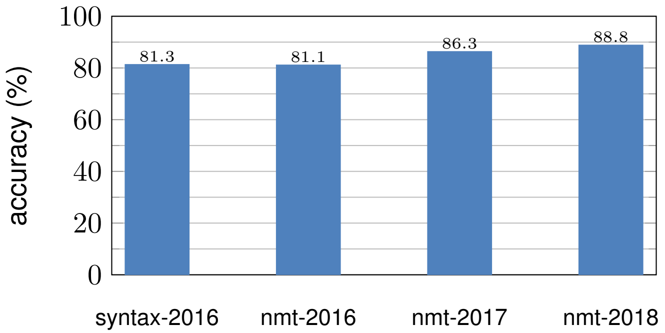


ContraWSD test set

- 35 ambiguous German nouns
- 2–4 senses per source noun
- $\approx$ 100 test instances per sense
  $\rightarrow \approx$ 7000 test instances
- ways to evaluate:
  - is reference more probable than contrastive variant?
  - does translation contain correct sense, wrong sense, or both/neither?

source:	*Also nahm ich meinen amerikanischen Reisepass und stellte mich in die **Schlange** für Extranjeros.*
reference:	*So I took my U.S. passport and got in the **line** for Extranjeros.*
contrastive:	*So I took my U.S. passport and got in the **snake** for Extranjeros.*
contrastive:	*So I took my U.S. passport and got in the **serpent** for Extranjeros.*

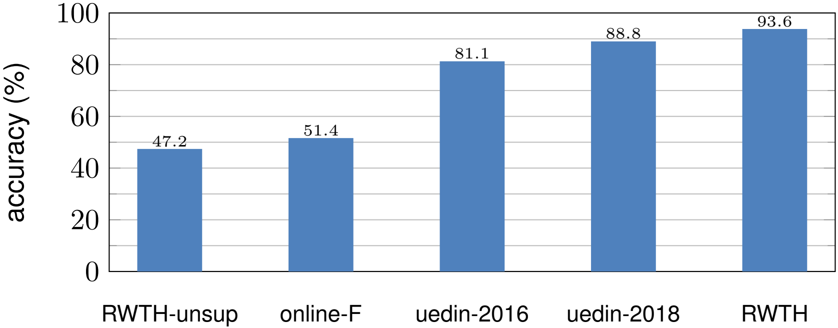
ContraWSD Results (uedin systems)



improvements to NMT systems

- 2016: shallow RNN
- 2017: deep RNN; layer norm; better ensembles; slightly more data
- 2018: Transformer; more (noisy) data

ContraWSD Results (selected systems)



- WSD is big challenge for unsupervised NMT and rule-based system
- all neural systems at WMT18 > 81%
- big reduction in WSD errors within 2 years

So... Time to Retire?

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

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

March 14, 2018 | Allison Linn

SDL Cracks Russian to English Neural Machine Translation

Global Enterprises to Capitalize on Near Perfect Russian to English Machine Translation as SDL Sets New Industry Standard

June 19, 2018, Maidenhead, UK

Coreference and Consistency
[Läubli, Sennrich, Volk, EMNLP 2018]


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. [...]

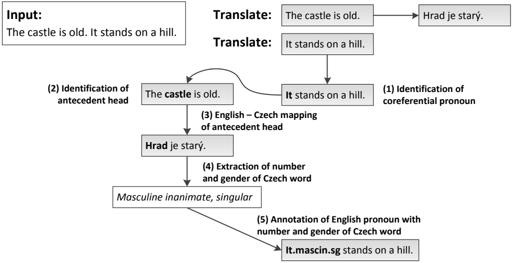
What We Need to Go Beyond Sentence Level

Models
make prediction conditional on context beyond the sentence

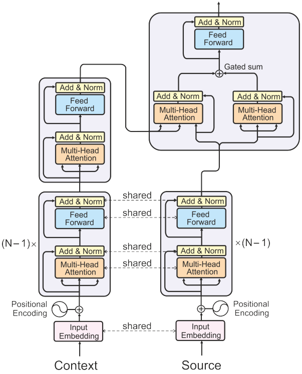
Metrics
measure improvements in consistency, and on less-frequent phenomena

Data
provide full document pairs as training data / deal with lack thereof

Models for Context-Aware MT



context-aware SMT architecture

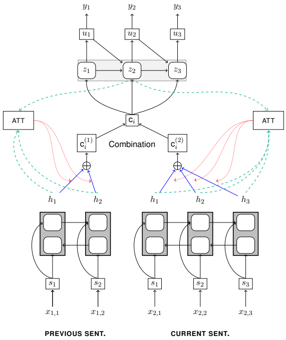


context-aware NMT architecture

[Guillou, 2012, Voita et al., 2018]

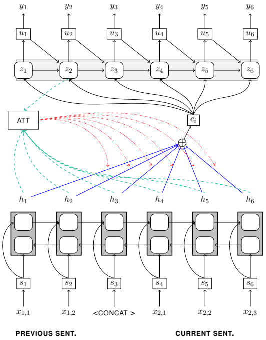
Models for Context-Aware NMT

multi-source architectures



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

concatenation strategy



[Tiedemann and Scherrer, 2017]

[Bawden et al., 2018]

Context-Aware Transformer Learns Anaphora Resolution

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



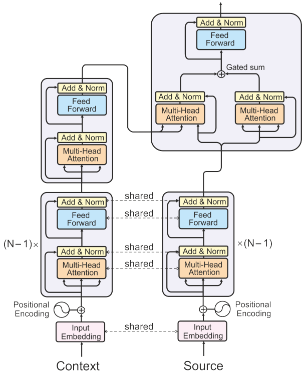
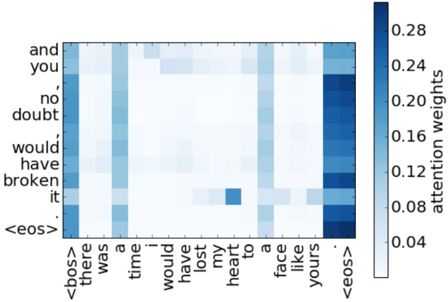


Figure 1: Encoder of the discourse-aware model



	agreement
coreNLP	77%
attention	72%
last noun	54%

Agreement with human assessment for coreference resolution of anaphoric *it*.

Metrics for Context-Aware MT

problems with standard metrics (BLEU etc.)

- local
- reference-based (not measuring consistency) [Guillou and Hardmeier, 2018]
- appropriate for long tail?

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?

Contrastive Evaluation

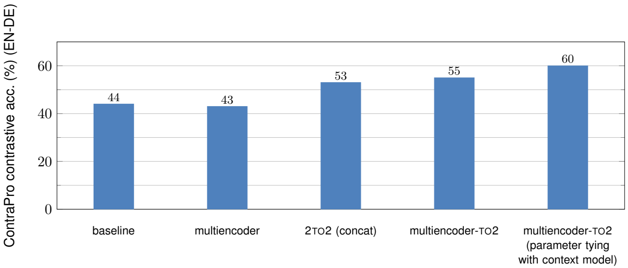
[Bawden, Sennrich, Birch, Haddow, NAACL 2018]
[Müller, Rios, Voita, Sennrich, WMT 2018]
[Voita, Sennrich, Titov, ACL 2019]

test sets targeting phenomena such as:

- anaphoric pronouns
- consistency in formality (T-V distinction)
- consistency in named entity translation
- translation of elliptical constructions

reference is paired with **contrastive variants** that introduce error
$\rightarrow$ we count how often MT system prefers correct variant

Some Lessons From Contrastive Evaluation



[Müller et al., 2018]

- even simple concatenation models bring substantial improvements
- small design decisions matter:
  learning context model from scratch suboptimal
- difficulty varies across linguistic phenomena

(Lack of) Data for Context-Aware MT

30 years of data collection in MT: **sentence pairs**

can we shift to document-level parallel corpora?

- requires extra work and reprocessing for some corpora
- impossible for others
  (e.g. bitext mining from comparable corpora)

Using Monolingual Document-level Data

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]

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

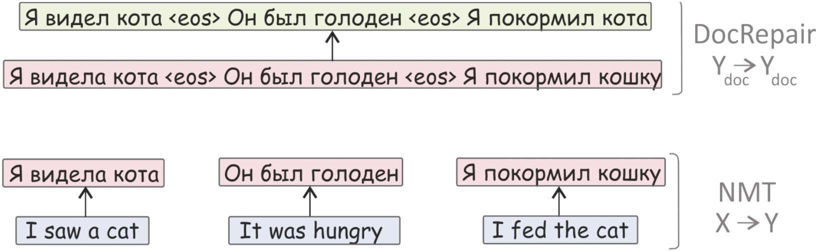
- channel model ($P(S|T)$) operates on sentence-level.
- language model ($P(T)$) operates on document-level.

solution 2: automatic post-editing (monolingual repair)

Context-Aware Monolingual Repair

[Voita, Sennrich, Titov, EMNLP 2019]

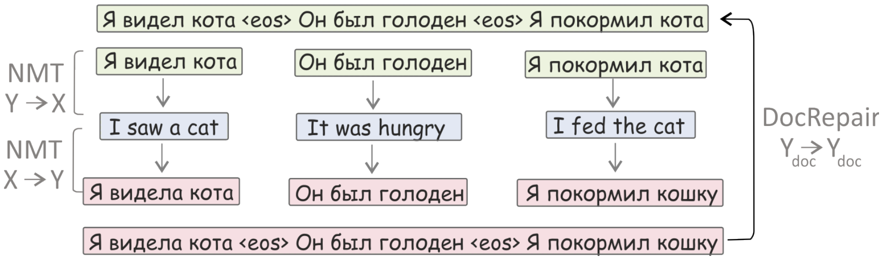
1. translate sentences independently
2. fix inconsistencies with multi-sentence monolingual repair model



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



Results on Consistency Test Sets [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**
monolingual repair	**34.6**	**91.8**	**80.6**	**86.4**	75.2

Conclusions

- neural MT models strong at learning from context
- current challenge: going beyond the sentence level
  - better metrics for development and measuring progress
    $\rightarrow$ small design decisions have big impact on ”context-awareness“!
  - document-level datasets...
    ...and models that work without document-level parallel data

Thank you for your attention

Resources

- ContraWSD test set for Word Sense Disambiguation:
  https://github.com/ZurichNLP/ContraWSD
- 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

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 645452 (QT21), HimL (644402), and ELITR (825460). Further funding from the European Research Council was received from the ERC Starting Grant BroadSem (678254).

This work has received funding from the Swiss National Science Foundation (SNF) in the projects CoNTra (grant number 105212_169888) and the Synergia MODERN project (grant number 147653).

This work has received funding from the Dutch National Science Foundation (NWO VIDI639.022.518).

This work has received funding from the Royal Society (NAF\R1\180122).

Bibliography I

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.

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., 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.

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.

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

Rios, A., Mascarell, L., and Sennrich, R. (2017).
Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings.
In Proceedings of the Second Conference on Machine Translation, Volume 1: Research Papers, Copenhagen, Denmark.

Rios, A., Müller, M., and Sennrich, R. (2018).
The Word Sense Disambiguation Test Suite at WMT18.
In Proceedings of the Third Conference on Machine Translation, pages 594–602, Belgium, Brussels.

Sennrich, R. and Haddow, B. (2016).
Linguistic Input Features Improve Neural Machine Translation.
In Proceedings of the First Conference on Machine Translation, Volume 1: Research Papers, pages 83–91, Berlin, Germany.

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.

Bibliography III

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.

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.

Image Credits

- racket: https://www.flickr.com/photos/128067141@N07/15157111178 / CC BY 2.0
- attacker: https://commons.wikimedia.org/wiki/File:Wikibully.jpg
- bat1: www.personalcreations.com / CC-BY-2.0
- bat2: Hasitha Tudugalle https://commons.wikimedia.org/wiki/File:Flying-Fox-Bat.jpg / CC-BY-4.0