How Contextual is Neural Machine Translation?

Rico Sennrich
a caricature of rule-based MT:
a caricature of rule-based MT:

- let’s translate a sentence word-by-word via a bilingual dictionary 😊
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... 😊
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 😊
Er hat einen Krebtest 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 😊
Er has entwickelt einen Krebstest

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
How Contextual is Neural Machine Translation?

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

<table>
<thead>
<tr>
<th>system</th>
<th>sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>source reference</td>
<td>Dort wurde er von dem <strong>Schläger</strong> und einer weiteren männl. Person erneut angegriffen. There he was attacked again by his <strong>original attacker</strong> and another male.</td>
</tr>
<tr>
<td>our NMT Google</td>
<td>There he was attacked again by the <strong>racket</strong> and another male person. There he was again attacked by the <strong>bat</strong> and another male person.</td>
</tr>
<tr>
<td>system</td>
<td>sentence</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>source reference</td>
<td>Dort wurde er von dem <strong>Schläger</strong> und einer weiteren männl. Person erneut angegriffen. There he was attacked again by his <strong>original attacker</strong> and another male.</td>
</tr>
<tr>
<td>our NMT Google</td>
<td>There he was attacked again by the <strong>racket</strong> and another male person.</td>
</tr>
<tr>
<td></td>
<td>There he was again attacked by the <strong>bat</strong> and another male person.</td>
</tr>
<tr>
<td>system</td>
<td>sentence</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>source</td>
<td>Dort wurde er von dem <strong>Schläger</strong> und einer weiteren männl. Person erneut angegriffen.</td>
</tr>
<tr>
<td>reference</td>
<td></td>
</tr>
<tr>
<td>our NMT</td>
<td>There he was attacked again by the <strong>racket</strong> and another male person.</td>
</tr>
<tr>
<td>Google</td>
<td></td>
</tr>
</tbody>
</table>

**Schläger**

- **racket**
- attacker; thug
<table>
<thead>
<tr>
<th>system</th>
<th>sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>Dort wurde er von dem <strong>Schläger</strong> und einer weiteren männl. Person erneut angegriffen. There he was attacked again by his <strong>original attacker</strong> and another male.</td>
</tr>
<tr>
<td>reference</td>
<td>There he was attacked again by his <strong>original attacker</strong> and another male.</td>
</tr>
<tr>
<td>our NMT</td>
<td>There he was attacked again by the <strong>racket</strong> and another male person.</td>
</tr>
<tr>
<td>Google</td>
<td>There he was again attacked by the <strong>bat</strong> and another male person.</td>
</tr>
</tbody>
</table>

**Schläger**

- **racket**
- **attacker; thug**
- **bat**
Dort wurde er von dem **Schläger** und einer weiteren männl. Person erneut angegriffen. There he was attacked again by his **original attacker** and another male.

<table>
<thead>
<tr>
<th>system</th>
<th>sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>source reference</td>
<td>Dort wurde er von dem <strong>Schläger</strong> und einer weiteren männl. Person erneut angegriffen. There he was attacked again by his <strong>original attacker</strong> and another male.</td>
</tr>
<tr>
<td>our NMT Google</td>
<td>There he was attacked again by the <strong>racket</strong> and another male person. There he was again attacked by the <strong>bat</strong> and another male person.</td>
</tr>
</tbody>
</table>

**Schläger**

racket  attacker; thug  bat
We thought a win like this might be close. Wir dachten, dass ein solcher Sieg nah sein könnte. *Wir dachten, ein Sieg wie dieser könnte schließen.
Generic Knowledge and Word Sense Disambiguation

Etymology 1
From Old English clýsan (“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 claudó.

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

\[
E_1(\text{close}) = \begin{bmatrix} 0.4 \\ 0.1 \\ 0.2 \end{bmatrix} \quad E_2(\text{adj}) = \begin{bmatrix} 0.1 \end{bmatrix}
\]

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

source reference
NMT (uedin WMT16) +POS, dependency, lemma, morphology

We thought a win like this might be close_adj. Wir dachten, dass ein solcher Sieg nah sein könnte.
Wir dachten, ein Sieg wie dieser könnte schließen. Wir dachten, ein Sieg wie dieser könnte nah sein.
ContraWSD test set

- 35 ambiguous German nouns
- 2–4 senses per source noun
- \( \approx 100 \) test instances per sense
  \[ \rightarrow \approx 7000 \text{ test instances} \]
- 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. |
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
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
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. [...]
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

Input: The castle is old. It stands on a hill.

Translate: The castle is old. Hrad je starý.

Context-aware SMT architecture

![Diagram of SMT architecture](image)

context-aware NMT architecture

![Diagram of NMT architecture](image)

[Guillou, 2012, Voita et al., 2018]
multi-source architectures

concatenation strategy

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

[Tiedemann and Scherrer, 2017]

[Bawden et al., 2018]
Table 7: Agreement with human assessment for coreference resolution of anaphoric *it*.

<table>
<thead>
<tr>
<th>Agreement</th>
<th>coreNLP</th>
<th>77%</th>
</tr>
</thead>
<tbody>
<tr>
<td>attention</td>
<td>72%</td>
<td></td>
</tr>
<tr>
<td>last noun</td>
<td>54%</td>
<td></td>
</tr>
</tbody>
</table>

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

Figure 1: Encoder of the discourse-aware model
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

In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously unexplored valley in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English. The scientists named the population, after their distinctive horn, What’s Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Paredes, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Paredes noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Paredes and the others then ventured further into the valley. “By the time we reached the top of one peak, the water looked blue, with some crystals on top,” said Paredes. Paredes and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them. If they were so close they could touch their horns.

While examining these bizarre creatures, the scientists discovered that the creatures also spoke some fairly regular English. Paredes stated, “We can see, for example, that they have a common language, something like a dialect or dialectic.”

Dr. Paredes believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America. While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Paredes, “In South America, such incidents seem to be quite common.” However, Paredes also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. “But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization,” said the scientist.

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

can we distinguish accidental repetition from document-level cohesion?
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 → we count how often MT system prefers correct variant
Some Lessons From Contrastive Evaluation

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

[Müller et al., 2018]
(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)
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)**
translate sentences independently

fix inconsistencies with multi-sentence monolingual repair model

1. sample several groups of sentences from the monolingual data;
2. for each sentence in a group, (i) translate it using a target-to-source MT model, (ii) sample a translation of this back-translated sentence in the source language using a source-to-target MT model;
3. using these round-trip translations of isolated sentences, form an inconsistent version of the initial groups;
4. use inconsistent groups as input for the DocRepair model, consistent ones as output.

At test time, the process of getting document-level translations is two-step (Figure 2):
1. produce translations of isolated sentences using a context-agnostic MT model;
2. apply the DocRepair model to a sequence of context-agnostic translations to correct inconsistencies between translations.

In the scope of the current work, the DocRepair model is the standard sequence-to-sequence Transformer. Sentences in a group are concatenated using a reserved token-separator between sentences.

The Transformer is trained to correct these long inconsistent pseudo-sentences into consistent ones. The token-separator is then removed from corrected translations.
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

Our key contributions are as follows:

• we introduce the first approach to context-aware machine translation using only monolingual document-level data;
• our approach shows substantial improvements in translation quality as measured by BLEU, targeted contrastive evaluation of several discourse phenomena and human evaluation;
• we show which discourse phenomena are hard to capture using monolingual data only.
## Results on Consistency Test Sets

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

<table>
<thead>
<tr>
<th>system</th>
<th>BLEU</th>
<th>deixis</th>
<th>lexical cohesion</th>
<th>ellipsis (infl.)</th>
<th>ellipsis (VP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentence-level</td>
<td>33.9</td>
<td>50.0</td>
<td>45.9</td>
<td>53.0</td>
<td>28.4</td>
</tr>
<tr>
<td>concatenation (4-to-4)</td>
<td>-</td>
<td>83.5</td>
<td>47.5</td>
<td>76.2</td>
<td><strong>76.6</strong></td>
</tr>
<tr>
<td>monolingual repair</td>
<td><strong>34.6</strong></td>
<td><strong>91.8</strong></td>
<td><strong>80.6</strong></td>
<td><strong>86.4</strong></td>
<td>75.2</td>
</tr>
</tbody>
</table>
Conclusions

- neural MT models strong at learning from context
- current challenge: going beyond the sentence level
  - better metrics for development and measuring progress
    → 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
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).
Evaluating Discourse Phenomena in Neural Machine Translation.
In NAACL 2018, New Orleans, USA.

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 Computational Linguistics, pages 1–10, Avignon, France.

Automatic reference-based evaluation of pronoun translation misses the point.

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.

In EMNLP 2018, Brussels, Belgium.

A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation.
Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings.

The Word Sense Disambiguation Test Suite at WMT18.

Linguistic Input Features Improve Neural Machine Translation.

Neural Machine Translation with Extended Context.
In Proceedings of the Third Workshop on Discourse in Machine Translation, pages 82–92, Copenhagen, Denmark.


When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion.
In ACL 2018, Melbourne, Australia.

Exploiting Cross-Sentence Context for Neural Machine Translation.

Extending machine translation evaluation metrics with lexical cohesion to document level.

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