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

March 14, 2018 | Allison Linn

- 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
can we reproduce Microsoft’s finding with different evaluation protocol?

<table>
<thead>
<tr>
<th></th>
<th>original evaluation</th>
<th>our evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>test set</td>
<td>WMT17</td>
<td>WMT17 (native Chinese part)</td>
</tr>
<tr>
<td>system</td>
<td>Microsoft COMBO-6</td>
<td>Microsoft COMBO-6</td>
</tr>
<tr>
<td>raters</td>
<td>crowd-workers</td>
<td>professional translators</td>
</tr>
<tr>
<td>experimental unit</td>
<td>sentence</td>
<td>sentence / document</td>
</tr>
<tr>
<td>measurement</td>
<td>direct assessment</td>
<td>pairwise ranking</td>
</tr>
<tr>
<td>raters see reference</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>raters see source ratings</td>
<td>yes</td>
<td>yes / no</td>
</tr>
<tr>
<td>ratings</td>
<td>≥ 2,520 per system</td>
<td>≈ 200 per setting</td>
</tr>
</tbody>
</table>
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 a vehicle and is unable to contact the owner of the obstructing vehicle can use the "WeChat Move the Car" function to address the issue.
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. [...]
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. [...]

Rico Sennrich
Evaluation Results: Adequacy Assessment

![Bar chart showing the winner in pairwise ranking (in %) for MT, TIE, and HUMAN. The chart indicates that HUMAN wins with 41.0%, followed by MT with 50.0%, and TIE with 9.0%.

Legend:
- Blue bars represent sentence-level
- Red bars represent document-level

Rico Sennrich  Document-level Neural Machine Translation  4/42
Evaluation Results: Adequacy Assessment

![Bar chart showing winner in pairwise ranking (in %) for MT, TIE, and HUMAN at sentence-level and document-level.]

- MT: 50.0% winner at sentence-level, 37.0% at document-level
- TIE: 9.0% winner at sentence-level, 11.0% at document-level
- HUMAN: 41.0% winner at sentence-level, 52.0% at document-level

Legend:
- Sentence-level
- Document-level
Evaluation Results: Fluency Assessment

![Graph showing fluency assessment results for MT, TIE, and HUMAN in pairwise ranking (in %).]

- MT: 32.0%
- TIE: 17.0%
- HUMAN: 51.0%

Legend: blue for sentence-level, red for document-level.
Evaluation Results: Fluency Assessment

![Graph showing winner in pairwise ranking (in %)]

- **MT**:
  - Sentence-level: 32.0%
  - Document-level: 22.0%

- **TIE**:
  - Sentence-level: 17.0%
  - Document-level: 29.0%

- **HUMAN**:
  - Sentence-level: 51.0%
  - Document-level: 50.0%
Follow-Up Study: Error Analysis

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

![Error Analysis Diagram]

- Incorrect word: MT 85, Human 51
- Missing word: MT 56, Human 37
- Named entity: MT 30, Human 16
- Word order: MT 17, Human 1
- Collocation: MT 27, Human 15
- Context (register/coref): MT 12, Human 6

(sentences containing error of this type (N=150))
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
  → WMT 2019 is shifting to document-level evaluation
SMT era:

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

**Translate:**
The castle is old. Hrad je starý.

**Translate:**
It stands on a hill.

Figure 1: Overview of the Annotation Process

- **(1) Identification of coreferential pronoun**
  - **Input:** The castle is old. It stands on a hill.
  - **Translate:**
    - The castle is old. Hrad je starý.
    - It stands on a hill.

- **(2) Identification of antecedent head**
  - **Input:** The castle is old. It stands on a hill.
  - **Translate:**
    - The castle is old.
    - It stands on a hill.

- **(3) English – Czech mapping of antecedent head**
  - **Input:** It stands on a hill.
  - **Translate:**
    - Hrad je starý.

- **(4) Extraction of number and gender of Czech word**
  - **Input:** Hrad je starý.
  - **Translate:**
    - Hrad je starý.

- **(5) Annotation of English pronoun with number and gender of Czech word**
  - **Input:** It.masc.sg stands on a hill.
  - **Translate:**
    - It.masc.sg stands on a hill.

**specialized features**

New Chances: Context-Aware NMT

NMT era:

\[ t = ( \text{Erweiterter, Kontext, in, Neuronaler, Maschinel, Übersetzung} ) \]

\[ s = ( \text{Rück, Contexts, in, Neural, Machine, Translation} ) \]

contextual sentences as additional input

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?
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?

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.

Previous linguistic context necessary to disambiguate
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.
icorrect: T'inquiète, je le tuerai pour toi.

correct: T'inquiète, je les moucherons. Regarde, il y en a un autre !
icorrect: T'inquiète, je le tuerai pour toi.
icorrect: T'inquiète, je la tuerai pour toi.

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

Previous linguistic context necessary to disambiguate

Balanced examples:
Non-contextual baseline scores 50%
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.
icorrect: 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.
icorrect: C'est un peu plus cher que ce que je pensais.
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

2TO2 - concatenated input

Multiple encoders

[CBahdanau et al., 2015]

[Tiedemann and Scherrer, 2017]

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

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>19.5</td>
</tr>
<tr>
<td>concat-2T01</td>
<td>19.5</td>
</tr>
<tr>
<td>concat-2T02</td>
<td>20.1</td>
</tr>
<tr>
<td>multiencoder (source; hierarchical attention)</td>
<td>20.2</td>
</tr>
<tr>
<td>multiencoder-T02 (hierarchical attention)</td>
<td>20.9</td>
</tr>
</tbody>
</table>
Results: Contrastive Test Set: Coreference

- Baseline: 50.0%
- concat-2\rightarrow1: 52.0%
- concat-2\rightarrow2: 63.5%
- multiencoder (source; hierarchical attention): 50.0%
- multiencoder-\rightarrow2 (hierarchical attention): 72.5%
Results: Contrastive Test Set: Coherence/Cohesion

<table>
<thead>
<tr>
<th></th>
<th>accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>50.0</td>
</tr>
<tr>
<td>concat-2to1</td>
<td>53.0</td>
</tr>
<tr>
<td>concat-2to2</td>
<td>52.0</td>
</tr>
<tr>
<td>multiencoder (source; hierarchical attention)</td>
<td>53.0</td>
</tr>
<tr>
<td>multiencoder-to2 (hierarchical attention)</td>
<td>57.0</td>
</tr>
</tbody>
</table>
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

**BLEU (newstest2017)**

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>23.0</td>
</tr>
<tr>
<td>multiencoder</td>
<td>23.5</td>
</tr>
<tr>
<td>2TO2 (concat)</td>
<td>23.8</td>
</tr>
<tr>
<td>multiencoder-TO2</td>
<td>23.8</td>
</tr>
<tr>
<td>multiencoder-TO2 (with encoder embedding tying)</td>
<td>24.2</td>
</tr>
</tbody>
</table>

**ContraPro contrastive acc. (%)**

<table>
<thead>
<tr>
<th>Model</th>
<th>ContraPro acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>44</td>
</tr>
<tr>
<td>multiencoder</td>
<td>43</td>
</tr>
<tr>
<td>2TO2 (concat)</td>
<td>53</td>
</tr>
<tr>
<td>multiencoder-TO2</td>
<td>55</td>
</tr>
<tr>
<td>multiencoder-TO2 (with encoder embedding tying)</td>
<td>60</td>
</tr>
</tbody>
</table>
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?
multiencoder-TO2 has context window of 1:

- why does quality improve when nominal antecedent is in same sentence, or further away?
  → coreference chains
- why does baseline improve with increased antecedent distance?
multiencoder-TO2 has context window of 1:

- why does quality improve when nominal antecedent is in same sentence, or further away? → coreference chains
- why does baseline improve with increased antecedent distance? → more instances of majority class
Example with antecedent distance 2:

<table>
<thead>
<tr>
<th>source (EN)</th>
<th>t-2</th>
<th>t-1</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>What’s with <strong>the door</strong>?</td>
<td><strong>It</strong> won’t open.</td>
<td>- Is <strong>it</strong> locked?</td>
<td></td>
</tr>
<tr>
<td>target (DE)</td>
<td>Was ist mit <strong>der Tür</strong>?</td>
<td><strong>Sie</strong> geht nicht auf.</td>
<td>- Ist <strong>sie</strong> abgeschlossen?</td>
</tr>
</tbody>
</table>
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]
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

<table>
<thead>
<tr>
<th>one/both bad</th>
<th>both good bad pair</th>
<th>both good good pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>211</td>
<td>140</td>
<td>1649</td>
</tr>
<tr>
<td>11%</td>
<td>7%</td>
<td>82%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>type of error</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>deixis</td>
<td>37%</td>
</tr>
<tr>
<td>ellipsis</td>
<td>29%</td>
</tr>
<tr>
<td>lexical cohesion</td>
<td>14%</td>
</tr>
<tr>
<td>ambiguity</td>
<td>9%</td>
</tr>
<tr>
<td>anaphora</td>
<td>6%</td>
</tr>
<tr>
<td>other</td>
<td>5%</td>
</tr>
<tr>
<td>type of error</td>
<td>frequency</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>T-V distinction</td>
<td>67%</td>
</tr>
<tr>
<td>speaker/addressee gender:</td>
<td></td>
</tr>
<tr>
<td>same speaker</td>
<td>22%</td>
</tr>
<tr>
<td>different speaker</td>
<td>9%</td>
</tr>
<tr>
<td>other</td>
<td>2%</td>
</tr>
</tbody>
</table>

translation errors caused by deixis (excluding anaphora)
Deixis

<table>
<thead>
<tr>
<th>type of error</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-V distinction</td>
<td>67%</td>
</tr>
<tr>
<td>speaker/addressee gender:</td>
<td></td>
</tr>
<tr>
<td>same speaker</td>
<td>22%</td>
</tr>
<tr>
<td>different speaker</td>
<td>9%</td>
</tr>
<tr>
<td>other</td>
<td>2%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>type of error</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-V distinction</td>
<td>67%</td>
</tr>
<tr>
<td><strong>speaker/addressee gender:</strong></td>
<td></td>
</tr>
<tr>
<td>same speaker</td>
<td>22%</td>
</tr>
<tr>
<td>different speaker</td>
<td>9%</td>
</tr>
<tr>
<td>other</td>
<td>2%</td>
</tr>
</tbody>
</table>

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. Я пришла в Саймону не ради тебя. Я сделал это для себя.

feminine, masculine.
### Ellipsis

<table>
<thead>
<tr>
<th>type of error</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>wrong morphological form</td>
<td>66%</td>
</tr>
<tr>
<td>wrong verb (VP-ellipsis)</td>
<td>20%</td>
</tr>
<tr>
<td>other error</td>
<td>14%</td>
</tr>
</tbody>
</table>

*translation errors caused by ellipsis*
You call her your friend but have you been to her home? Her work?

Ty называешь её своей подругой, но ты был у неё дома? Её работа?

Wrong morphological form: noun phrase marked as subject
### Ellipsis

<table>
<thead>
<tr>
<th>type of error</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>wrong morphological form</td>
<td>66%</td>
</tr>
<tr>
<td>wrong verb (VP-ellipsis)</td>
<td>20%</td>
</tr>
<tr>
<td>other error</td>
<td>14%</td>
</tr>
</tbody>
</table>

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.

---

**En** Not for Julia. Julia has a taste for taunting her victims.

**Ru** Не для Джулии. Юлия умеет дразнить своих жертв.

---

Inconsistent translation

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

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 scientist named the population, after their distinctive horn, Unicorns. 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 Párredez, 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. Párredez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Párredez 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 Párredez.

Párredez 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. Párredez stated, "We can see, for example, that they have a common language, something like a dialect or dialectic."

Dr. Párredez 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 Párredez, "In South America, such incidents seem to be quite common."

However, Párredez 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.

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

<table>
<thead>
<tr>
<th></th>
<th>latest relevant context</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total</td>
</tr>
<tr>
<td>deixis</td>
<td>3000</td>
</tr>
<tr>
<td>lexical cohesion</td>
<td>2000</td>
</tr>
<tr>
<td>ellipsis (inflection)</td>
<td>500</td>
</tr>
<tr>
<td>ellipsis (VP)</td>
<td>500</td>
</tr>
</tbody>
</table>

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 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

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>32.4</td>
</tr>
<tr>
<td>concatenate</td>
<td>31.6</td>
</tr>
<tr>
<td>CADec (ours)</td>
<td>32.4</td>
</tr>
</tbody>
</table>
Contrastive Results

**deixis test set**
- Baseline: 50.0
- Concatenate: 83.5
- CADec (ours): 81.6

**lexical cohesion test set**
- Baseline: 45.9
- Concatenate: 47.5
- CADec (ours): 58.1
Contrastive Results

Ellipsis Test Set (Inflection)

- Baseline: 53.0%
- Concatenate: 76.2%
- CADec (ours): 72.2%

Ellipsis Test Set (VP)

- Baseline: 28.4%
- Concatenate: 76.6%
- CADec (ours): 80.0%
### Results: Choice of First-Pass Translation During Training

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

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!**
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
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:
  → 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
Contextual Handling in Neural Machine Translation: Look Behind, Ahead and on Both Sides.
In 21th Annual Conference of the European Association for Machine Translation.

Neural Machine Translation by Jointly Learning to Align and Translate.

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.

Discourse in statistical machine translation: A survey and a case study.
Discours, 11.

Context-Aware Learning for Neural Machine Translation.
CoRR, abs/1903.04715.

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.

Document Context Neural Machine Translation with Memory Networks.

Machine Translation of Labeled Discourse Connectives.
In Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA).

A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in 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.

In ACL 2018, Melbourne, Australia.

Exploiting Cross-Sentence Context for Neural Machine Translation.
Extending machine translation evaluation metrics with lexical cohesion to document level.
Figure 6: Attention patterns with referential pronouns in extended context.
Analyzing Use of Context: Transformer

The animal didn’t cross the street because it was too tired.

The animal didn’t cross the street because it was too wide.

The animal didn’t cross the street because it was too tired.

The animal didn’t cross the street because it was too wide.
set-up

- Transformer architecture with clear interface to context
- analysis of attention patterns
The self-attention mechanism first computes attention weights: i.e., for each word, it computes a distribution over all words (including itself). This distribution is then used to compute a new representation of that word: this new representation is set to an expectation (under the attention distribution specific to the word) of word representations from the layer below. In multi-head attention, this process is repeated \( h \) times with different representations and the result is concatenated.

The second component of each layer of the Transformer network is a feed-forward network. The authors propose using a two-layered network with the ReLU activations.

Analogously, each layer of the decoder contains the two sub-layers mentioned above as well as an additional multi-head attention sub-layer that receives input from the corresponding encoding layer.

In the decoder, the attention is masked to prevent future positions from being attended to, or in other words, to prevent illegal leftward information flow. See Vaswani et al. (2017) for additional details.

The proposed architecture reportedly improves over the previous best results on the WMT 2014 English-to-German and English-to-French translation tasks, and we verified its strong performance on our data set in preliminary experiments. Thus, we consider it a strong state-of-the-art baseline for our experiments. Moreover, as the Transformer is attractive in practical NMT applications because of its parallelizability and training efficiency, integrating extra-sentential information in Transformer is important from the engineering perspective. As we will see in Section 4, previous techniques developed for recurrent encoder-decoders do not appear effective for the Transformer.

3 Context-aware model architecture

Our model is based on Transformer architecture (Vaswani et al., 2017). We leave Transformer's decoder intact while incorporating context information on the encoder side (Figure 1).

**Source encoder:**
The encoder is composed of a stack of \( N \) layers. The first \( N-1 \) layers are identical and represent the original layers of Transformer's encoder. The last layer incorporates contextual information as shown in Figure 1. In addition to multi-head self-attention it has a block which performs multi-head attention over the output of the context encoder stack. The outputs of the two attention mechanisms are combined via a gated sum. More precisely, let \( c(s\text{-attn})_i \) be the output of the multi-head self-attention, \( c(c\text{-attn})_i \) the output of the multi-head attention to context, \( c_i \) their gated sum, and \( \sigma \) the logistic sigmoid function, then

\[
g_i = \sigma(W_g[c(s\text{-attn})_i, c(c\text{-attn})_i] + b_g)
\]

\[
c_i = g_i \odot c(s\text{-attn})_i + (1 - g_i) \odot c(c\text{-attn})_i
\]

**Context encoder:**
The context encoder is composed of a stack of \( N \) identical layers and replicates the original Transformer encoder. In contrast to related work (Jean et al., 2017; Wang et al., 2017), we found in preliminary experiments that using separate encoders does not yield an accurate model. Instead we share the parameters of the first \( N-1 \) layers with the source encoder. Since major proportion of the context encoder's parameters are shared with the source encoder, we add a special token (let us denote it \(<\text{bos}>\) ) to the beginning of context sentences, but not source

---

Figure 1: Encoder of the discourse-aware model
Context-Aware Transformer: Evaluation

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

![Graph showing BLEU scores]

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

![Graph showing BLEU scores]

Rico Sennrich
Table 7: Agreement with CoreNLP for test sets of pronouns having a nominal antecedent in context sentence (%). Examples with \( \geq 1 \) noun in context sentence. Note that an agreement of the last noun for "it" or the first noun for "you" and "I" is very high. This is partially due to the fact that most context sentences have only one noun. For these examples a random and last predictions are always correct, meanwhile attention does not always pick a noun as the most relevant word in the context. To get a more clear picture let us now concentrate only on examples where there is more than one noun in the context (Table 7). We can now see that the attention weights are in much better agreement with the coreference system than any of the heuristics. This indicates that the model is indeed performing anaphora resolution.

While agreement with CoreNLP is encouraging, we are aware that coreference resolution by CoreNLP is imperfect and partial agreement with it may not necessarily indicate that the attention is particularly accurate. In order to control for this, we asked human annotators to manually evaluate 500 examples from the test sets where CoreNLP predicted that "it" refers to a noun in the context sentence. More precisely, we picked random 500 examples from the test set with "it" from Table 7. We marked the pronoun in a source which CoreNLP found anaphoric. Assessors were given the source and context sentences and were asked to mark an antecedent noun phrase for a marked pronoun in a source sentence or say that there is no antecedent at all. We then picked those examples where assessors found a link from "it" to some noun in context (79% of all examples). Then we evaluated agreement of CoreNLP and our model with the ground truth links. We also report the performance of the best heuristic for "it" from our previous analysis (i.e. last noun in context). The results are provided in Table 8.

The agreement between our model and the ground truth is 72\%. Though 5% below the coreference system, this is a lot higher than the best agreement (in %) CoreNLP 77\% attention 72\% last noun 54\% Table 8: Performance of CoreNLP and our model's attention mechanism compared to human assessment. Examples with \( \geq 1 \) noun in context sentence.

Rico Sennrich

6 Related work

Our analysis focuses on how our context-aware neural model implicitly captures anaphora. Early work on anaphora phenomena in statistical machine translation has relied on external systems for coreference resolution (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010). Results

Figure 5: An example of an attention map between source and context. On the y-axis are the source tokens, on the x-axis the context tokens. Note the high attention between "it" and its antecedent "heart".

CoreNLP right wrong
attn right 53 19
attn wrong 24 4

Table 9: Performance of CoreNLP and our model's attention mechanism compared to human assessment (%). Examples with \( \geq 1 \) noun in context sentence.

heuristic (+18%). This confirms our conclusion that our model performs latent anaphora resolution. Interestingly, the patterns of mistakes are quite different for CoreNLP and our model (Table 9). We also present one example (Figure 5) where the attention correctly predicts anaphora while CoreNLP fails. Nevertheless, there is room for improvement, and improving the attention component is likely to boost translation performance.
<table>
<thead>
<tr>
<th></th>
<th>agreement (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>coreNLP</td>
<td>77</td>
</tr>
<tr>
<td>attention</td>
<td>72</td>
</tr>
<tr>
<td>last noun</td>
<td>54</td>
</tr>
</tbody>
</table>

Agreement with human assessment for coreference resolution of anaphoric *it*. Examples with $\geq 1$ noun in context sentence.
In fairness, Miller did not attack the statue itself. 
[...] 
But he did attack its meaning [...]
In fairness, Miller did not attack the statue itself.  
[...]  
But he did attack its meaning [...]  

<table>
<thead>
<tr>
<th>HUMAN</th>
<th>MT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Um fair zu bleiben, Miller griff nicht die Statue selbst an.</td>
<td>Fairerweise hat Miller die Statue nicht selbst angegriffen.</td>
</tr>
<tr>
<td>[...]</td>
<td>[...]</td>
</tr>
<tr>
<td>Aber er griff deren Bedeutung an [...]</td>
<td>Aber er griff seine Bedeutung an [...]</td>
</tr>
</tbody>
</table>
### Examples from Top WMT18 Systems

**Lexical Coherence**

<table>
<thead>
<tr>
<th>HUMAN</th>
<th>MT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pasture fence project is fundamental</td>
<td>Electric fence project is basic</td>
</tr>
<tr>
<td>The Fischerbach pasture fence project is a successful project and will be continued next year.</td>
<td>The Fischerbacher Weidezaun-Project is a success and will be continued in the coming year.</td>
</tr>
</tbody>
</table>
### Examples from Top WMT18 Systems

#### lexical coherence

**HUMAN**

<table>
<thead>
<tr>
<th>Pasture fence project is fundamental</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Fischerbach pasture fence project is a successful project and will be continued next year.</td>
</tr>
</tbody>
</table>

**MT**

<table>
<thead>
<tr>
<th>Electric fence project is basic</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Fischerbacher Weidezaun-Project is a success and will be continued in the coming year.</td>
</tr>
</tbody>
</table>
### 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）[...]

<table>
<thead>
<tr>
<th><strong>HUMAN</strong></th>
<th><strong>MT</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>This robot uses speech synthesis, [...] with conversational [...] features.</td>
<td>Using speech synthesis [...] the robot has the functions of chatting conversation [...]</td>
</tr>
<tr>
<td>It has won two major CES awards [...]</td>
<td>Has won two awards at the International Consumer Electronics Exhibition (CES) [...]</td>
</tr>
</tbody>
</table>
This robot uses speech synthesis, [...] with conversational [...] features.

It has won two major CES awards [...]