CUBBITT
Charles University Block-Backtranslation-Improved Transformer Translation

Mgr. Martin Popel, Ph.D.
ÚFAL (Institute of Formal and Applied Linguistics), MFF UK

https://lindat.cz/cubbitt
Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar & Zdeněk Žabokrtský

Nature Communications 11, Article number: 4381 (2020) | Cite this article

6273 Accesses | 76 Altmetric | Metrics
History of the research

- professors Sgall, Hajičová, Panevová, Hajič, Kuboň...
  2007 my interest in CL & NLP
professors Sgall, Hajičová, Panevová, Hajič, Kuboň...
2007 my interest in CL & NLP
2012/02 Markéta&Jakub use NN for t-lemma translation
History of the research

- professors Sgall, Hajičová, Panevová, Hajič, Kuboň...
  2007 my interest in CL & NLP
- 2012/02 Markéta & Jakub use NN for t-lemma translation
- 2017/01 I use DeepNN for Doctors Without Borders
History of the research

- professors Sgall, Hajičová, Panevová, Hajič, Kuboň...
  2007 my interest in CL & NLP
- 2012/02 Markéta&Jakub use NN for t-lemma translation
- 2017/01 I use DeepNN for Doctors Without Borders
- 2017/06 Attention Is All You Need (TectoMT→Transformer)
- 2018/04 Training Tips for the Transformer Model
- 2018/07 WMT18: CUBBIT signif. better than human ref.
- 2018/07 finished my PhD thesis
- 2018/08 new doc-aware evaluations, contacting Łukasz
- 2018/12 submission to Science, 2019/02 reviews, /08 reject
- 2019/09 submission to NatCommun, 2020/02 reviews, /04 resubmit, 07/accepted, 09/published
professors Sgall, Hajičová, Panevová, Hajič, Kuboň…
2007 my interest in CL & NLP
2012/02 Markéta&Jakub use NN for t-lemma translation
2017/01 I use DeepNN for Doctors Without Borders
2017/06 Attention Is All You Need (TectoMT→Transformer)
2018/04 Training Tips for the Transformer Model
History of the research

- professors Sgall, Hajičová, Panevová, Hajič, Kuboň...
- 2007 my interest in CL & NLP
- 2012/02 Markéta&Jakub use NN for t-lemma translation
- 2017/01 I use DeepNN for Doctors Without Borders
- 2017/06 Attention Is All You Need (TectoMT→Transformer)
- 2018/04 Training Tips for the Transformer Model
- 2018/07 WMT18: CUBBITT signif. better than human ref.
professors Sgall, Hajičová, Panevová, Hajič, Kuboň...
2007 my interest in CL & NLP
2012/02 Markéta & Jakub use NN for t-lemma translation
2017/01 I use DeepNN for Doctors Without Borders
2017/06 Attention Is All You Need (TectoMT $\rightarrow$ Transformer)
2018/04 Training Tips for the Transformer Model
2018/07 WMT18: CUBBITT signif. better than human ref.
2018/07 finished my PhD thesis
History of the research

- professors Sgall, Hajičová, Panevová, Hajič, Kuboň...
  2007 my interest in CL & NLP
- 2012/02 Markéta&Jakub use NN for t-lemma translation
- 2017/01 I use DeepNN for Doctors Without Borders
- 2017/06 Attention Is All You Need (TectoMT → Transformer)
- 2018/04 Training Tips for the Transformer Model
- 2018/07 WMT18: CUBBITT signif. better than human ref.
- 2018/07 finished my PhD thesis
- 2018/08 new doc-aware evaluations, contacting Łukasz
 professors Sgall, Hajičová, Panevová, Hajič, Kuboň...
2007 my interest in CL & NLP
2012/02 Markéta&Jakub use NN for t-lemma translation
2017/01 I use DeepNN for Doctors Without Borders
2017/06 Attention Is All You Need (TectoMT\rightarrow Transformer)
2018/04 Training Tips for the Transformer Model
2018/07 WMT18: CUBBITT signif. better than human ref.
2018/07 finished my PhD thesis
2018/08 new doc-aware evaluations, contacting Łukasz
2018/12 submission to Science, 2019/02 reviews, /08 reject
History of the research

- professors Sgall, Hajičová, Panevová, Hajič, Kuboň...
  - 2007 my interest in CL & NLP
- 2012/02 Markéta&Jakub use NN for t-lemma translation
- 2017/01 I use DeepNN for Doctors Without Borders
- 2017/06 Attention Is All You Need (TectoMT→Transformer)
- 2018/04 Training Tips for the Transformer Model
- 2018/07 WMT18: CUBBITT signif. better than human ref.
- 2018/07 finished my PhD thesis
- 2018/08 new doc-aware evaluations, contacting Łukasz
- 2018/12 submission to Science, 2019/02 reviews, /08 reject
- 2019/09 submission to NatCommun, 2020/02 reviews, experiments, /04 resubmit, 07/accepted, 09/published
The main result of our evaluation

56% sentences translated more adequately by CUBBITT, 34% sentences by a professional translation agency.
The main innovation: Block-BackTranslation

![Graph showing BLEU scores over training time for different methods: block-BT, block-BT + avg8, mix-BT, mix-BT + avg8. The graph illustrates how the BLEU scores evolve over time, highlighting the improvements achieved by Block-BackTranslation.]
Manual evaluation example

blind, sentence-level but document-aware, side-by-side (RankME)

<table>
<thead>
<tr>
<th>Source</th>
<th>Translation1</th>
<th>T1: overall</th>
<th>T1: judicious</th>
<th>Translation2</th>
<th>T2: overall</th>
<th>T2: judicious</th>
<th>Optional comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;And we’re protecting our shareholders from employment tightness.&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Companies started taking ethics, values and employee engagement more seriously in 2002 after accounting firm Arthur Andersen collapsed because of ethical violations from the Enron scandal, Quinlan said.&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;But it wasn’t until &quot;social media came into its own&quot; that companies realized they couldn’t stop their dirty laundry from going viral online.&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Prior to using technology to monitor ethics, people used hope as a strategy,&quot; he said.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;It’s a business model and opportunity that has the approval of venture capital investors, who have pumped up both start-ups.&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convergent raised $10 million in funding in February from firms such as Sapphire Ventures and Tola Capital, bringing its total capital raised to $47 million.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glint secured $10 million in November from Bessemer Venture Partners, bringing its total funding to $60 million.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>These investments hardly come as a surprise, given the interconnected nature of companies, culture and venture capital.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>There’s a growing body of research showing today’s employees expect more from their workplaces than before.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In competitive markets such as Silicon Valley, high salaries and interesting projects are merely table stakes.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employees want to feel that they’re accepted and valued and that they’re giving their time to a company with a positive mission.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
It depends on the input text domain

<table>
<thead>
<tr>
<th>Domain</th>
<th>n</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business</td>
<td>45</td>
<td>6.1e-04***</td>
</tr>
<tr>
<td>Crime</td>
<td>90</td>
<td>6.4e-04***</td>
</tr>
<tr>
<td>Entertainment</td>
<td>49</td>
<td>0.011*</td>
</tr>
<tr>
<td>Politics</td>
<td>115</td>
<td>0.003**</td>
</tr>
<tr>
<td>Scitech</td>
<td>65</td>
<td>0.410</td>
</tr>
<tr>
<td>Sport</td>
<td>90</td>
<td>0.011*</td>
</tr>
<tr>
<td>World</td>
<td>58</td>
<td>0.092</td>
</tr>
</tbody>
</table>

- C

Sentence-level adequacy - CUBBITT - Reference

- Median
- Mean
It depends on the input text domain.
Are three measures necessary?

Can we predict the overall quality as a weighted average of adequacy and fluency?
It depends on the annotator

Translators put more emphasis on fluency, non-professionals on adequacy.
Some annotators put all the emphasis on adequacy, but only for one of the systems.
We cannot conclude that professionals are more reliable.
CUBBITT makes more errors than humans in
- translation of ambiguous words
- fluency (but not spelling and grammar) and
- cross-sentence context

A

Context-aware evaluation of error types: 3 non-professionals and 3 professional translators
BlockBT improves the quality (+0.4 adequacy, +0.3 fluency).
60% of participants did not distinguish CUBBITT from human translations (on 100 isolated sentences)
BLEU Training curves for en-fr and en-pl
Why Block-BT works?

A

Source sentence: "He was an original guy and lived life to the full" said Gray in a statement.

SYNTH: chlap  AUTH: člověk  SYNTH: Šedivý  AUTH: Gray

Translation of block-BT-Avg: "Byl to originální chlap a žil život naplno," uvedl Gray v prohlášení.

Checkpoint 1 (SYNTH): "Byl to originální chlap a žil život naplno" uvedl Šedivý v prohlášení.
Checkpoint 2 (SYNTH): "Byl to originální chlap a žil život naplno" uvedl Šedivý v prohlášení.
Checkpoint 3 (AUTH): „Byl to originální člověk a žil život naplno,“ uvedl Gray v prohlášení.
Checkpoint 4 (AUTH): „Byl to originální člověk a žil život naplno,“ uvedl Gray v prohlášení.
Checkpoint 5 (AUTH): „Byl to originální člověk a žil život naplno,“ uvedl Gray v prohlášení.
Checkpoint 6 (AUTH): „Byl to originální člověk a žil život naplno,“ uvedl Gray v prohlášení.
Checkpoint 7 (AUTH): „Byl to originální člověk a žil život naplno,“ uvedl Gray v prohlášení.
Checkpoint 8 (AUTH): „Byl to originální člověk a žil život naplno,“ uvedl Gray v prohlášení.

B

![Graph showing relative probability of better translation over training time](image)