Revisiting Challenges in Neural Machine Translation

Rico Sennrich

University of Edinburgh
Why Revisit Challenges Regularly?

- guide research directions
- NMT facts have expiration date

signpost: Ian Harding (CC BY-NC-SA 2.0)
Some Challenges in Neural MT

1. Long Sentences
2. Adequacy
3. Low-Resource Translation

Challenges Revisited

1. Long Sentences
2. Adequacy
3. Low-Resource Translation

Future Challenges
Encoder–Decoder Has Information Bottleneck

[Cho et al., 2014]
Attention Brings Improvement

[Bahdanau et al., 2015]
Still, Poor Performance Reported for Long Sentences

Figure 7: Quality of translations based on sentence length. SMT outperforms NMT for sentences longer than 60 subword tokens. For very long sentences (80+) quality is much worse due to too short output.

3.5 Word Alignment

The key contribution of the attention model in neural machine translation (Bahdanau et al., 2015) was the imposition of an alignment of the output words to the input words. This takes the shape of a probability distribution over the input words which is used to weigh them in a bag-of-words representation of the input sentence.

Arguably, this attention model does not functionally play the role of a word alignment between the source in the target, at least not in the same way as its analog in statistical machine translation. While in both cases, alignment is a latent variable that is used to obtain probability distributions over words or phrases, arguably the attention model has a broader role. For instance, when translating a verb, attention may also be paid to its subject and object since these may disambiguate it. To further complicate matters, the word representations are products of bidirectional gated recurrent neural networks that have the effect that each word representation is informed by the entire sentence context.

But there is a clear need for an alignment mechanism between source and target words. For instance, prior work used the alignments provided by the attention model to interpolate word translation decisions with traditional probabilistic dictionaries (Arthur et al., 2016), for the introduction of coverage and fertility models (Tu et al., 2016), etc. But is the attention model in fact the proper relations between Obama and Netanyahu have been strained for years. Die Beziehungen zwischen Obama und Netanyahu sind seit Jahren angespannt.
Some Challenges in Neural MT

- Long Sentences
- Adequacy
- Low-Resource Translation

Challenges Revisited

- Long Sentences
- Adequacy
- Low-Resource Translation

Future Challenges
Adequacy vs. Fluency in WMT16 Evaluation

Adequacy
+1%

Fluency
+13%

Figure: WMT16 direct assessment results
comparison of NMT and PBSMT for EN→{DE, EL, PT, RU}

direct assessment:
- NMT obtains higher fluency judgment than PBSMT: +10%
- NMT only obtains small improvement in adequacy judgment: +1%

### Error Annotation

<table>
<thead>
<tr>
<th>category</th>
<th>SMT</th>
<th>NMT</th>
<th>difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>inflectional morphology</td>
<td>2274</td>
<td>1799</td>
<td>-21%</td>
</tr>
<tr>
<td>word order</td>
<td>1098</td>
<td>691</td>
<td>-37%</td>
</tr>
<tr>
<td>omission</td>
<td>421</td>
<td>362</td>
<td>-14%</td>
</tr>
<tr>
<td>addition</td>
<td>314</td>
<td>265</td>
<td>-16%</td>
</tr>
<tr>
<td>mistranslation</td>
<td>1593</td>
<td>1552</td>
<td>-3%</td>
</tr>
<tr>
<td>&quot;no issue&quot;</td>
<td>449</td>
<td>788</td>
<td>+75%</td>
</tr>
</tbody>
</table>
Some Challenges in Neural MT
- Long Sentences
- Adequacy
- Low-Resource Translation

Challenges Revisited
- Long Sentences
- Adequacy
- Low-Resource Translation

Future Challenges
Low-Resource Translation

Figure 3: BLEU scores for English-Spanish systems.

Figure 4: Translations of the first sentence of the data.

[Reference: Koehn and Knowles, 2017]
Some Challenges in Neural MT
- Long Sentences
- Adequacy
- Low-Resource Translation

Challenges Revisited
- Long Sentences
- Adequacy
- Low-Resource Translation

Future Challenges
Why Are Long Sentences Hard?

different answers

- training on long sentences efficiently is challenging
  → training–test mismatch
- locally normalized models have bias towards low-entropy states
  → outputs too short (</s>)
- long-distance interactions may be challenging due to network path length (vanishing gradient)
- ...?
Long Sentences: Training–Test Mismatch

Figure 7: Quality of translations based on sentence length. SMT outperforms NMT for sentences longer than 60 subword tokens. For very long sentences (80+) quality is much worse due to too short output.

3.5 Word Alignment

The key contribution of the attention model in neural machine translation (Bahdanau et al., 2015) was the imposition of an alignment of the output words to the input words. This takes the shape of a probability distribution over the input words which is used to weigh them in a bag-of-words representation of the input sentence.

Arguably, this attention model does not functionally play the role of a word alignment between the source in the target, at least not in the same way as its analog in statistical machine translation. While in both cases, alignment is a latent variable that is used to obtain probability distributions over words or phrases, arguably the attention model has a broader role. For instance, when translating a verb, attention may also be paid to its subject and object since these may disambiguate it. To further complicate matters, the word representations are products of bidirectional gated recurrent neural networks that have the effect that each word representation is informed by the entire sentence context.

But there is a clear need for an alignment mechanism between source and target words. For instance, prior work used the alignments provided by the attention model to interpolate word translation decisions with traditional probabilistic dictionaries (Arthur et al., 2016), for the introduction of coverage and fertility models (Tu et al., 2016), etc.

But is the attention model in fact the proper relations between Obama and Netanyahu have been strained for years. die Beziehungen zwischen Obama und Netanjahu sind seit Jahren angespannt.

Figure 8: Word alignment for English–German: comparing the attention model states (green boxes with probability in percent if over 10) with alignments obtained from fast-align (blue outlines).

Means? To examine this, we compare the soft alignment matrix (the sequence of attention vectors) with word alignments obtained by traditional word alignment methods. We use incremental fast-align (Dyer et al., 2013) to align the input and output of the neural machine system. See Figure 8 for an illustration. We compare the word attention states (green boxes) with the word alignments obtained with fast align (blue outlines). For most words, these match up pretty well. Both attention states and fast-align alignment points are a bit fuzzy around the function words have-been/sind. However, the attention model may settle on alignments that do not correspond with our intuition or alignment points obtained with fast-align.

See Figure 9 for the reverse language direction, German–English. All the alignment points appear to be off by one position. We are not aware of any intuitive explanation for this divergent behavior — the translation quality is high for both systems.

We measure how well the soft alignment (attention model) of the NMT system match the alignments of fast-align with two metrics:

- a match score that checks for each output if the aligned input word according to fast-align is indeed the input word that received the highest attention probability, and
- a probability mass score that sums up the probability mass of the input word that received the highest attention probability.

this is uedin-2016 system, trained with a maximum length of 50 subwords!
How to Train on Long Sentences

Problem: time and memory increases with longest sentence in batch

Solutions

- sort sentences of same length together [Sutskever et al., 2014]
- adjust batch size depending on length [Johansen et al., 2016]

<table>
<thead>
<tr>
<th>layer no.</th>
<th>units</th>
</tr>
</thead>
<tbody>
<tr>
<td>input alphabet size (X)</td>
<td>300</td>
</tr>
<tr>
<td>embedding sizes</td>
<td>256</td>
</tr>
<tr>
<td>char RNN (forward)</td>
<td>400</td>
</tr>
<tr>
<td>char RNN (backward)</td>
<td>400</td>
</tr>
<tr>
<td>attention</td>
<td>300</td>
</tr>
<tr>
<td>char decoder</td>
<td>400</td>
</tr>
</tbody>
</table>

Table 1. Hyperparameter values used for training the char-to-char model. Where $\Sigma_{\text{src}}$ and $\Sigma_{\text{trg}}$ represent the number of classes in the source and target languages, respectively.

<table>
<thead>
<tr>
<th>layer no.</th>
<th>units</th>
</tr>
</thead>
<tbody>
<tr>
<td>input alphabet size (X)</td>
<td>300</td>
</tr>
<tr>
<td>embedding sizes</td>
<td>256</td>
</tr>
<tr>
<td>char RNN (forward)</td>
<td>400</td>
</tr>
<tr>
<td>spaces RNN (forward)</td>
<td>400</td>
</tr>
<tr>
<td>spaces RNN (backward)</td>
<td>400</td>
</tr>
<tr>
<td>attention</td>
<td>300</td>
</tr>
<tr>
<td>char decoder</td>
<td>400</td>
</tr>
</tbody>
</table>

Table 2. Hyperparameter values used for training the char2word-to-char model. Where $\Sigma_{\text{src}}$ and $\Sigma_{\text{trg}}$ represent the number of classes in the source and target languages, respectively.

5.3. Results

5.3.1. Quantitative

The quantitative results of our models are illustrated in table 3. Notice that the char2word-to-char model outperforms the char-to-char model on all datasets (average 1.28 BLEU performance increase). This could be an indication that either having hierarchical, word-like, representations on the encoder or simply the fact that the encoder was significantly smaller, helps in NMT when using a character decoder with attention.
Training on Long Sentences Matters

![Graph](image)

**BLEU score by source sentence length (number of subwords)**

- **X-axis**: Source sentence length
- **Y-axis**: BLEU score

Graph showing the BLEU score for different source sentence lengths, comparing two conditions: max len 50 and max len 200.
Long Sentences: Label Bias

locally normalized models have label bias [Murray and Chiang, 2018]

→ (tunable) length penalty and variants result in simple globally normalized model
→ other methods to escape local normalization include reconstruction [Tu et al., 2017]
does network architecture affect learning of long-distance dependencies?

architectures

- RNN/GRU/LSTM
- convolution [Gehring et al., 2017]
- self-attention [Vaswani et al., 2017]

**Figure 1:** The Transformer - model architecture.

Decoder: The decoder is also composed of a stack of $N = 6$ identical layers. In addition to the two sub-layers in each encoder layer, the decoder inserts a third sub-layer, which performs multi-head attention over the output of the encoder stack. Similar to the encoder, we employ residual connections around each of the sub-layers, followed by layer normalization. We also modify the self-attention sub-layer in the decoder stack to prevent positions from attending to subsequent positions. This masking, combined with fact that the output embeddings are offset by one position, ensures that the predictions for position $i$ can depend only on the known outputs at positions less than $i$.

**3.2 Attention**

An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.

**3.2.1 Scaled Dot-Product Attention**

We call our particular attention “Scaled Dot-Product Attention” (Figure 2). The input consists of queries and keys of dimension $d_k$, and values of dimension $d_v$. We compute the dot products of the query with all keys, divide each by $\sqrt{d_k}$, and apply a softmax function to obtain the weights on the values.
### Evaluation with Contrastive Pairs: LingEval97 [Sennrich, EACL 2017]

<table>
<thead>
<tr>
<th></th>
<th>Sentence</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>[...] that the <strong>plan will</strong> be approved</td>
<td></td>
</tr>
<tr>
<td>German (correct)</td>
<td>[...] , dass der <strong>Plan</strong> verabschiedet <strong>wird</strong></td>
<td>0.1</td>
</tr>
<tr>
<td>German (contrastive)</td>
<td>* [...] , dass der <strong>Plan</strong> verabschiedet <strong>werden</strong></td>
<td>0.01</td>
</tr>
</tbody>
</table>

Subject-verb agreement
EN→DE WMT systems trained with Nematus

targeted evaluation of subject-verb agreement with Lingeval97

- GRU/LSTM much more stable than RNN for long distances
EN→DE WMT systems trained with Nematus

targeted evaluation of subject-verb agreement with Lingeval97

GRU/LSTM much more stable than RNN for long distances
- EN→DE WMT systems trained with Sockeye
- targeted evaluation of subject-verb agreement with Lingeval97

Accuracy vs Distance for different architectures:
- RNNS2S
- RNN-bideep
- ConvS2S
- Transformer

No evidence that Transformer or ConvS2S outperform LSTM for long-distance interactions.
EN→DE WMT systems trained with Sockeye
- targeted evaluation of subject-verb agreement with Lingeval97

no evidence that Transformer or ConvS2S outperform LSTM for long-distance interactions
strongest evidence for weakness of NMT on long sentences comes from old systems

discarding long sentences no longer necessary in NMT training

BLEU does not tell us *why* a system performs poorly on long sentences

- are translations too short?
  → train on long sentences; use global scores
- is grammaticality poor?
  → architectures matter, but long-distance interactions modelled well by GRU/LSTM and Transformer
Some Challenges in Neural MT
- Long Sentences
- Adequacy
- Low-Resource Translation

Challenges Revisited
- Long Sentences
- Adequacy
- Low-Resource Translation

Future Challenges
Targeted Analysis: Word Sense Disambiguation

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

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

**Diagram:**

- Schläger
- attacker

- Image of a person holding a bat
- Image of a person in a red shirt

---

Rico Sennrich

Revisiting Challenges in NMT 23 / 49
<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

racket https://www.flickr.com/photos/128067141@N07/15157111178 / CC BY 2.0

attacker hasitha tudugalle https://commons.wikimedia.org/wiki/File:Flying-Fox-Bat.jpg / CC-BY-4.0
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.

There he was attacked again by the racket and another male person. There he was again attacked by the bat and another male person.
<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.</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](https://www.flickr.com/photos/128067141@N07/1515711178 / CC BY 2.0)

[attacker](https://commons.wikimedia.org/wiki/File:Flying-Fox-Bat.jpg / CC-BY -4.0)

[bat](https://commons.wikimedia.org/wiki/File:Flying-Fox-Bat.jpg / CC-BY -4.0)
test set (ContraWSD)

- 35 ambiguous German nouns
- 2–4 senses per source noun
- contrastive translation sets (1 or more contrastive translations)
- \( \approx 100 \) test instances per sense
  \( \rightarrow \approx 7000 \) test instances

---

**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.
Word Sense Accuracy

WSD is challenging, especially for rare word senses.
WSD is challenging, especially for rare word senses.
Word Sense Disambiguation: Measuring Progress
[Rios, Müller, Sennrich, WMT 2018]

based on ContraWSD, but semi-automatic evaluation of 1-best output
evaluating all WMT 2018 submissions, plus systems from previous years
Results: Word Sense Disambiguation (uedin systems)

<table>
<thead>
<tr>
<th>Year</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>syntax-2016</td>
<td>81.3</td>
</tr>
<tr>
<td>nmt-2016</td>
<td>81.1</td>
</tr>
<tr>
<td>nmt-2017</td>
<td>86.3</td>
</tr>
<tr>
<td>nmt-2018</td>
<td>88.8</td>
</tr>
</tbody>
</table>

improvements to NMT system

- **2016**: shallow RNN
- **2017**: deep RNN; layer normalization; better ensembles; slightly more training data
- **2018**: Transformer; more (noisy) training data
WSD is big challenge for unsupervised NMT and rule-based system

all neural systems at WMT18 > 81%

big reduction in WSD errors in last 2 years
- comparing different architectures on same dataset
- Transformer no better than RNN at long-distance agreement
- interesting differences for word sense disambiguation:

![Graph showing BLEU and ContraWSD accuracy for RNN, CNN, and Transformer architectures.](image)
post-publication experiments:
models can be made more similar to Transformer with:
- multihead attention
- feedforward block
- layer normalization
- ...

Transformer still ahead in WSD accuracy
Some Challenges in Neural MT
- Long Sentences
- Adequacy
- Low-Resource Translation

Challenges Revisited
- Long Sentences
- Adequacy
- Low-Resource Translation

Future Challenges
Conventional wisdom states that neural machine translation model, we find that NMT systems actually outperform systems of similar quality for German–English translation. To illustrate this, see Figure 4. With the input, some key words are properly translated into the output. However, both NMT and SMT systems continue to have difficulty translating some words. However, both NMT and SMT systems do continue to have difficulty translating some words. However, both NMT and SMT systems do continue to have difficulty translating some words.

Estrategia siria para contrarrestar la reelección de Obama. To obtain a learning curve, we used all additionally provided monolingual data for a big language model in contrastive SMT systems.
aspects worth revisiting:

- phrase-based benefit from large LM but NMT can also improve with monolingual data
- there were general improvements in NMT do they move the point where NMT outperforms SMT?
- the NMT system was not optimized for low-resource NMT does tuning model to low-resource NMT help?
There is a large pool of methods to exploit monolingual data for NMT:

- **ensembling with LM** [Gülçehre et al., 2015]
- **training objective: language modelling** [Sennrich et al., 2016, Ramachandran et al., 2016]
- **training objective: autoencoders** [Luong et al., 2016, Currey et al., 2017]
- **training objective: round-trip translation** [Sennrich et al., 2016, He et al., 2016, Cheng et al., 2016]
- **unsupervised NMT** [Artetxe et al., 2017, Lample et al., 2017]

Similarly, parallel data from other language pairs can help

[Zoph et al., 2016, Chen et al., 2017, Nguyen and Chiang, 2017]
There is a large pool of methods to exploit monolingual data for NMT:

- ensembling with LM [Gülçehre et al., 2015]
- training objective: language modelling [Sennrich et al., 2016, Ramachandran et al., 2016]
- training objective: autoencoders [Luong et al., 2016, Currey et al., 2017]
- training objective: round-trip translation [Sennrich et al., 2016, He et al., 2016, Cheng et al., 2016]
- unsupervised NMT [Artetxe et al., 2017, Lample et al., 2017]

Similarly, parallel data from other language pairs can help [Zoph et al., 2016, Chen et al., 2017, Nguyen and Chiang, 2017]

..but how far can we get with just parallel data?
Low-Resource Translation: Experiments

**setup**

- IWSLT14 German → English:
  - full set: 160,000 sentences (3.2M words)
  - smallest subset: 5,000 sentences (100,000 words)
- phrase-based SMT with Moses
- neural MT with Nematus and BPE
- baseline: hyperparameters similar to uedin@WMT16 shallow RNN, no dropout

**comparison to [Koehn and Knowles, 2017]**

[Koehn and Knowles, 2017]:
0.4 million to 385 million words of data (EN → ES WMT)

our experiments:
0.1 million to 3.2 million words of data (DE → EN IWSLT)
Low-Resource Translation: Experiments

- general architecture improvements:
  - bideep RNN [Miceli Barone et al., 2017]
  - layer normalization [Ba et al., 2016]
  - label smoothing [Szegedy et al., 2016]

- choices optimized for low-resource scenario:
  - dropout [Srivastava et al., 2014]
  - tied embeddings [Press and Wolf, 2017]
  - smaller BPE vocabulary size for smaller data sets
  - smaller batch size for smaller data sets
  - lexical model [Nguyen and Chiang, 2018]
Low-Resource Translation: Results

[ Koehn and Knowles, 2017 ]

**BLEU Scores with Varying Amounts of Training Data**

For the neural machine translation model, we use a publicly available model settings of Edinburgh’s WMT submission (Sennrich et al., 2016a). This was trained using a big 2 billion word in-domain language model with a phrase-based SMT system trained on each subset, respectively. In addition to a neural MT baseline, we also compare the neural system to a phrase-based SMT system of similar quality for German–English translation models perform particularly poorly on rare words, (Luong et al., 2015; Sennrich et al., 2016b; Arthur et al., 2016) due in part to the highly-inflected categories. SMT systems continue to have difficulty translating some infrequent words, particularly those belonging to rare words, (Luong et al., 2015; Sennrich et al., 2016a). This was trained using a publicly available model (https://github.com/rsennrich/wmt16-scripts/).

**Explosión realiza una estrategia divisiva de luchar contra las elecciones de autor libre determinación de Obama.**

For the neural machine translation model, we use a publicly available model settings of Edinburgh’s WMT submission (Sennrich et al., 2016a). This was trained using a big 2 billion word in-domain language model with a phrase-based SMT system trained on each subset, respectively. In addition to a neural MT baseline, we also compare the neural system to a phrase-based SMT system of similar quality for German–English translation models perform particularly poorly on rare words, (Luong et al., 2015; Sennrich et al., 2016b; Arthur et al., 2016) due in part to the highly-inflected categories. SMT systems continue to have difficulty translating some infrequent words, particularly those belonging to rare words, (Luong et al., 2015; Sennrich et al., 2016a). This was trained using a publicly available model (https://github.com/rsennrich/wmt16-scripts/).
Low-Resource Translation: Results

[Coehn and Knowles, 2017]

BLEU Scores with Varying Amounts of Training Data

![Graph showing BLEU scores with varying amounts of training data.]

Our experiments

![Graph comparing performance of different translation models.]

Corpus Size (English Words)

Corpus size (English words)
Low-Resource Translation: Results

![Graph showing BLEU scores for various corpus sizes in English words.]

- **neural MT optimized**
- **phrase-based SMT**
- **neural MT baseline**

**Corpus Size (English words)**

- **BLEU**
  - 0
  - 0.8
  - 1.2
  - 10.6
  - 16
  - 20.6
  - 23.1
  - 24
  - 24.3
  - 26.9
  - 29.5

Rico Sennrich

Revisiting Challenges in NMT
## Low-Resource Translation: Results

<table>
<thead>
<tr>
<th>system</th>
<th>BLEU by corpus size (words)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100k</td>
</tr>
<tr>
<td></td>
<td>3.2M</td>
</tr>
<tr>
<td>phrase-based SMT</td>
<td>14.1</td>
</tr>
<tr>
<td>NMT baseline</td>
<td>0.0</td>
</tr>
<tr>
<td>+dropout, tied embeddings, layer normalization, bideep RNN, label smoothing</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td>30.0</td>
</tr>
<tr>
<td>+reduce BPE vocabulary (14k → 2k symbols)</td>
<td>9.3</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td>+reduce batch size (4k → 1k tokens)</td>
<td>9.7</td>
</tr>
<tr>
<td></td>
<td>29.9</td>
</tr>
<tr>
<td>+lexical model</td>
<td>11.0</td>
</tr>
<tr>
<td></td>
<td>29.5</td>
</tr>
</tbody>
</table>
the balance between PBSMT and NMT for low-resource settings is shifting with
- general improvements in NMT
- careful choice of hyperparameters and architectures for low-resource setting

it is no longer true that we cannot train NMT on less than 1M words...
...but low-resource machine translation remains a challenge
1 Some Challenges in Neural MT
   - Long Sentences
   - Adequacy
   - Low-Resource Translation

2 Challenges Revisited
   - Long Sentences
   - Adequacy
   - Low-Resource Translation

3 Future Challenges
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

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
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

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

...extraordinary claims require extraordinary evidence
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
can we reproduce Microsoft’s finding with different evaluation protocol?

<table>
<thead>
<tr>
<th>Original Evaluation</th>
<th>Our Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Set</td>
<td>WMT17</td>
</tr>
<tr>
<td>System</td>
<td>Microsoft COMBO-6</td>
</tr>
<tr>
<td>Raters</td>
<td>Crowd-workers</td>
</tr>
<tr>
<td>Experimental Unit</td>
<td>Sentence</td>
</tr>
<tr>
<td>Measurement</td>
<td>Direct Assessment</td>
</tr>
<tr>
<td>Raters see Reference</td>
<td>No</td>
</tr>
<tr>
<td>Raters see Source</td>
<td>Yes</td>
</tr>
<tr>
<td>Ratings</td>
<td>≥ 2,520 per system</td>
</tr>
<tr>
<td></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 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，在日常生活中，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. [...]
Evaluation Results: Bilingual Assessment

Winner in pairwise ranking (in %)

MT: 50.0%
TIE: 9.0%
HUMAN: 41.0%

(sentence-level: blue, document-level: red)
Evaluation Results: Bilingual Assessment

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

- **MT**: 50.0 (sentence-level), 37.0 (document-level)
- **TIE**: 9.0 (sentence-level), 11.0 (document-level)
- **HUMAN**: 41.0 (sentence-level), 52.0 (document-level)
Evaluation Results: Monolingual Assessment

Winner in pairwise ranking (in %)

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

Legend:
- **sentence-level**
- **document-level**
Evaluation Results: Monolingual Assessment

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

- **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%

Legend:
- Blue: sentence-level
- Red: document-level
A Case for Document-level Evaluation

- document-level ratings show significant preference for HUMAN
- preference for HUMAN is even stronger in monolingual evaluation

Conclusions

- distinguishing MT from human translations becomes harder with increasing quality
- document-level evaluation shows some limitations of current NMT systems
Conclusions

- NMT has made tremendous progress in past years
  → to make progress, we need to regularly re-evaluate its weaknesses
- word sense disambiguation, long sentences, low-resource settings are still challenging, but no longer embarassingly bad
- plenty of challenges remain, such as document-level translation
Thank you for your attention

Resources

- WSD Test Suite: https://github.com/a-rios/ContraWSD
- Evaluation data on human parity: https://github.com/laeubli/parity
Unsupervised Neural Machine Translation.  
CoRR, abs/1710.11041.

Layer Normalization.  
CoRR, abs/1607.06450.

Neural Machine Translation by Jointly Learning to Align and Translate.  

Evaluating MT for massive open online courses.  
Machine Translation.


Semi-Supervised Learning for Neural Machine Translation.  

Copied Monolingual Data Improves Low-Resource Neural Machine Translation.

Convolutional Sequence to Sequence Learning.
CoRR, abs/1705.03122.

CoRR, abs/1503.03535.

Dual Learning for Machine Translation.

Neural Machine Translation with Characters and Hierarchical Encoding.
CoRR, abs/1610.06550.
Six Challenges for Neural Machine Translation.

CoRR, abs/1711.00043.


Multi-task Sequence to Sequence Learning.
In ICLR 2016.

Deep Architectures for Neural Machine Translation.

Correcting Length Bias in Neural Machine Translation.
Improving Lexical Choice in Neural Machine Translation.


Using the Output Embedding to Improve Language Models.

Unsupervised pretraining for sequence to sequence learning.

Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings.

The Word Sense Disambiguation Test Suite at WMT18.


