Neural Machine Translation: Breaking through the Performance Ceiling

Rico Sennrich

University of Edinburgh; Universität Zürich

April 19 2018
Statistical Machine Translation (SMT)

given a sequence of words $s$ in the source language, find the most probable sequence $t$ in the target language [Brown et al., 1993]

$$t^* \approx \arg \max_t \sum_{m=1}^{M} \lambda_m h_m(s, t)$$ [Och, 2003]
Statistical Machine Translation (SMT)

given a sequence of words \( s \) in the source language, find the most probable sequence \( t \) in the target language \[ \text{[Brown et al., 1993]} \]

\[ t^* \approx \arg \max_t \sum_{m=1}^{M} \lambda_m h_m(s, t) \] \[ \text{[Och, 2003]} \]

main research trend

feature engineering
Statistical Machine Translation (SMT)

given a sequence of words $s$ in the source language, find the most probable sequence $t$ in the target language [Brown et al., 1993]

$$t^* \approx \arg\max_t \sum_{m=1}^{M} \lambda_m h_m(s, t)$$ [Och, 2003]

main research trend

feature engineering
Neural Machine Translation

\[ f = (\text{La, croissance, économique, s'est, ralentie, ces, dernières, années, .} \) \]

\[ e = (\text{Economic, growth, has, slowed, down, in, recent, years, .}) \]
very general model:
- applied to many sequence–to–sequence tasks
- variants used in computer vision
Neural Machine Translation: Timeline

1987 Early encoder-decoder, with vocabulary size 30-40 [Allen, 1987]

2013 Pure neural MT system presented [Kalchbrenner and Blunsom, 2013]

2014 RNN Encoder-Decoder with Attention [Bahdanau et al., 2015]

2015 WMT 15: Montreal NMT is competitive [Jean et al., 2015b]

2015 Subword-level NMT [Sennrich et al., 2016c]

2015 Monolingual Data in NMT [Sennrich et al., 2016b]

2016 WMT 16: Edinburgh NMT is dominant [Sennrich et al., 2016a]

2017 Various architectures competitive [Gehring et al., 2017, Vaswani et al., 2017]
(tied) best constrained system for **7 out of 8** translation directions

<table>
<thead>
<tr>
<th>system</th>
<th>BLEU</th>
<th>official rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>uedin-nmt</td>
<td>34.2</td>
<td>1</td>
</tr>
<tr>
<td>metamind</td>
<td>32.3</td>
<td>2</td>
</tr>
<tr>
<td>uedin-syntax</td>
<td>30.6</td>
<td>3</td>
</tr>
<tr>
<td>NYU-UMontreal</td>
<td>30.8</td>
<td>4</td>
</tr>
<tr>
<td>online-B</td>
<td>29.4</td>
<td>5-10</td>
</tr>
<tr>
<td>KIT/LIMSI</td>
<td>29.1</td>
<td>5-10</td>
</tr>
<tr>
<td>cambridge</td>
<td>30.6</td>
<td>5-10</td>
</tr>
<tr>
<td>online-A</td>
<td>29.9</td>
<td>5-10</td>
</tr>
<tr>
<td>prompt-rule</td>
<td>23.4</td>
<td>5-10</td>
</tr>
<tr>
<td>KIT</td>
<td>29.0</td>
<td>6-10</td>
</tr>
<tr>
<td>jhu-syntax</td>
<td>26.6</td>
<td>11-12</td>
</tr>
<tr>
<td>jhu-pbmt</td>
<td>28.3</td>
<td>11-12</td>
</tr>
<tr>
<td>uedin-pbmt</td>
<td>28.4</td>
<td>13-14</td>
</tr>
<tr>
<td>online-F</td>
<td>19.3</td>
<td>13-15</td>
</tr>
<tr>
<td>online-G</td>
<td>23.8</td>
<td>14-15</td>
</tr>
</tbody>
</table>

WMT16 EN→DE

<table>
<thead>
<tr>
<th>system</th>
<th>BLEU</th>
<th>official rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>uedin-nmt</td>
<td>38.6</td>
<td>1</td>
</tr>
<tr>
<td>online-B</td>
<td>35.0</td>
<td>2-5</td>
</tr>
<tr>
<td>online-A</td>
<td>32.8</td>
<td>2-5</td>
</tr>
<tr>
<td>uedin-syntax</td>
<td>34.4</td>
<td>2-5</td>
</tr>
<tr>
<td>KIT</td>
<td>33.9</td>
<td>2-6</td>
</tr>
<tr>
<td>uedin-pbmt</td>
<td>35.1</td>
<td>5-7</td>
</tr>
<tr>
<td>jhu-pbmt</td>
<td>34.5</td>
<td>6-7</td>
</tr>
<tr>
<td>online-G</td>
<td>30.1</td>
<td>8</td>
</tr>
<tr>
<td>jhu-syntax</td>
<td>31.0</td>
<td>9</td>
</tr>
<tr>
<td>online-F</td>
<td>20.2</td>
<td>10</td>
</tr>
</tbody>
</table>

WMT16 DE→EN
problem

word-level neural networks use one-hot encoding
→ closed and small vocabulary

this gets you 95% of the way...
... if you only care about automatic metrics
problem

word-level neural networks use one-hot encoding → closed and small vocabulary

divide gets you 95% of the way...
... if you only care about automatic metrics

why 95% is not enough

rare outcomes have high self-information

<table>
<thead>
<tr>
<th>source</th>
<th>The indoor temperature is very pleasant. Das Raumklima ist sehr angenehm. Die UNK ist sehr angenehm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>reference</td>
<td>[Bahdanau et al., 2015]</td>
</tr>
</tbody>
</table>

∴ X
Open-Vocabulary Neural MT

**problem**

word-level neural networks use one-hot encoding
→ closed and small vocabulary

this gets you 95% of the way...
... if you only care about automatic metrics

**why 95% is not enough**

rare outcomes have high self-information

| source | The *indoor temperature* is very pleasant.  
Das *Raumklima* ist sehr angenehm. |
| reference | [Bahdanau et al., 2015]  
[Jean et al., 2015a] |
| Die *UNK* ist sehr angenehm.  
Die *Innenpool* ist sehr angenehm. | [✗] [✗] |
Open-Vocabulary Neural MT

problem

word-level neural networks use one-hot encoding
→ closed and small vocabulary

this gets you 95% of the way...
... if you only care about automatic metrics

why 95% is not enough

rare outcomes have high self-information

source reference

The indoor temperature is very pleasant. Das Raumklima ist sehr angenehm.

[Bahdanau et al., 2015]

Die UNK ist sehr angenehm. Die Innenpool ist sehr angenehm.✓

[Jean et al., 2015a]

Die Innen+ temperatur ist sehr angenehm.✗

[Sennrich, Haddow, Birch, ACL 2016a]
goal

subword segmentation that:

- uses a closed vocabulary of subword units
- can represent open vocabulary (including unknown words)
- minimizes the sequence length (given the vocabulary size)

solution

- greedy compression algorithm: byte pair encoding (BPE) [Gage, 1994]
- we adapt BPE to word segmentation
- hyperparameter: vocabulary size

<table>
<thead>
<tr>
<th>vocabulary size</th>
<th>text</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>t+ h+ e i+ n+ d+ o+ o+ r t+ e+ m+ p+ e+ r+ a+ t+ u+ r+ e i+ s v+ e+ r+ y p+ l+ e+ a+ s+ a+ n+ t</td>
</tr>
<tr>
<td>1300</td>
<td>the in+ do+ or t+ em+ per+ at+ ure is very p+ le+ as+ ant</td>
</tr>
<tr>
<td>10300</td>
<td>the in+ door temper+ ature is very pleasant</td>
</tr>
<tr>
<td>50300</td>
<td>the indoor temperature is very pleasant</td>
</tr>
</tbody>
</table>
Subword NMT: Translation Quality

<table>
<thead>
<tr>
<th>BLEU</th>
<th>EN-DE</th>
<th>EN-RU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMT</td>
<td>24.4</td>
<td>24.3</td>
</tr>
<tr>
<td>word-level NMT (with back-off)</td>
<td>22.0</td>
<td>19.1</td>
</tr>
<tr>
<td>subword-level NMT</td>
<td>22.8</td>
<td>20.4</td>
</tr>
</tbody>
</table>

[Sennrich and Haddow, 2015, Haddow et al., 2015]
[Jean et al., 2015a]
Subword NMT: Translation Quality

NMT Results EN-RU

- subword-level
- word-level (with back-off)
- word-level (no back-off)

unigram $F_1$

training set frequency rank

50 000 500 000

50 000 500 000

50 000 500 000
why?
monolingual data
- is much less sparse than parallel data
- is useful for structured prediction
- may be used for domain adaptation

why is this hard?
- standard in SMT: monolingual LM as feature in linear model
- linear combination of NMT and LM barely effective [Gülçehre et al., 2015]

our solution
end-to-end training of NMT model with parallel and monolingual data
Monolingual Data in NMT

NMT is a conditional language model

\[ p(u_i) = f(z_i, u_{i-1}, c_i) \]

Problem

for monolingual training instances, source context \( c_i \) is missing
Monolingual Training Instances

solutions: missing data imputation for $c_i$

- missing data indicator: $\overrightarrow{0}$
  $\rightarrow$ works, but danger of catastrophic forgetting
- impute $c_i$ with neural network
  $\rightarrow$ we do this indirectly by back-translating the target sentence
Evaluation: English → German

**Bar Chart**

- ** syntax-based SMT: 24.4**
- **NMT parallel: 23.6**
- **+missing data indicator: 24.6**
- **+back-translation: 26.5**

**Legend:**
- Syntax-based SMT
- NMT parallel
- +missing data indicator
- +back-translation

**Source:** [Sennrich and Haddow, 2015]

**(NMT systems are ensemble of 4)**
Human Evaluation of Neural MT [Bojar et al., 2016]

Fluency
is translation good English?
+13%

Adequacy
is meaning preserved?
+1%

Figure: WMT16 direct assessment results
### Word Sense Disambiguation

<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
### Adequacy in Neural Machine Translation

#### Word Sense Disambiguation

<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 again attacked 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>

![Diagram](https://www.flickr.com/photos/128067141@N07/15157111178 / CC BY 2.0)

**Schläger**

There he was attacked again by his **original attacker** and another male.

**racket**

There he was again attacked by the **racket** and another male person.

**bat**

There he was again attacked by the **bat** and another male person.
Word Sense Disambiguation

<table>
<thead>
<tr>
<th>system</th>
<th>sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>source reference</td>
<td>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.</td>
</tr>
<tr>
<td>our NMT Google</td>
<td>There he was attacked again by the racket and another male person. There he was again attacked by the bat and another male person.</td>
</tr>
</tbody>
</table>
Adequacy in Neural Machine Translation

Word Sense Disambiguation

<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.</td>
</tr>
<tr>
<td>our NMT Google</td>
<td>There he was attacked again by his <strong>original attacker</strong> and another male.</td>
</tr>
<tr>
<td>Google</td>
<td>There he was again attacked 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>
</tbody>
</table>

Schläger

- **racket**
- **attacker**
- **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.

There he was attacked again by the racket and another male person. There he was again attacked by the bat and another male person.
core idea
- NMT assigns score to every translation hypothesis
- we provide NMT system with several translations:
  - correct human reference translation
  - contrastive variants which introduces error
- we count how often model prefers correct translation

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.
at WMT16, UEDIN-NMT was top-ranked
large lead in fluency; small lead in adequacy
for WMT17, we improved our MT system in several ways:
  - deep transition networks
  - layer normalization
  - better hyperparameters
  - better ensembles
  - (slightly) more training data

are we getting better at word sense disambiguation?
Results: Word Sense Disambiguation

accuracy (%)

word sense disambiguation accuracy
n=7359

UEDIN-NMT @ WMT16: single
UEDIN-NMT @ WMT17: single
UEDIN-NMT @ WMT17: ensemble
≈ human performance (sentence-level)
Results: Word Sense Disambiguation

Word sense disambiguation accuracy

n=7359

- UEDIN-NMT @ WMT16: single
- UEDIN-NMT @ WMT17: single
- UEDIN-NMT @ WMT17: ensemble
- ≈ human performance (sentence-level)
Results: Word Sense Disambiguation

word sense disambiguation accuracy
n=7359

- UEDIN-NMT @ WMT16: single
- UEDIN-NMT @ WMT17: single
- UEDIN-NMT @ WMT17: ensemble
- ≈ human performance (sentence-level)
Results: Word Sense Disambiguation

[n=7359]

- UEDIN-NMT @ WMT16: single
- UEDIN-NMT @ WMT17: single
- UEDIN-NMT @ WMT17: ensemble
- ≈ human performance (sentence-level)
What Did We Learn?

- Word sense disambiguation remains a challenging problem in MT, but measurable progress last year.
- On sentence-level, even humans may find it challenging.

German reference: "Sehen Sie die Muster?"
Contrastive: "Do you see the patterns?"
Contrastive: "Do you see the examples?"

→ Targeted evaluation of document-level modelling [Bawden et al., 2018]
Conclusion

neural sequence–to–sequence models

- neural models have revolutionized MT
- we have overcome early limitations
- many methods shared with general deep learning

open challenges

- increasing semantic faithfulness
  - scaling up sequence length (documents)
  - novel objective functions (NCE, GANs etc.)
  - word sense disambiguation
- data efficiency
  - interactive MT
  - one-shot learning
  - low-resourced translation
Collaborators

Alexandra Birch

Annette Rios

Anna Currey

Antonio Valerio Miceli Barone

Laura Mascarell

Ulrich Germann

Phil Williams

Barry Haddow

Martin Volk

Kenneth Heafield
Thank you for your attention

Resources

- **BPE scripts:** [https://github.com/rsennrich/subword-nmt](https://github.com/rsennrich/subword-nmt)
- **ContraWSD:** [https://github.com/a-rios/ContraWSD](https://github.com/a-rios/ContraWSD)
- **pre-trained models:**
  - **WMT16:** [http://data.statmt.org/wmt16_systems/](http://data.statmt.org/wmt16_systems/)
  - **WMT17:** [http://data.statmt.org/wmt17_systems/](http://data.statmt.org/wmt17_systems/)
Several Studies on Natural Language and Back-Propagation.
In IEEE First International Conference on Neural Networks, pages 335–341, San Diego, California, USA.

Neural Machine Translation by Jointly Learning to Align and Translate.

Evaluating Discourse Phenomena in Neural Machine Translation.
In NAACL 2018, New Orleans, USA.


Describing Multimedia Content using Attention-based Encoder-Decoder Networks.

A New Algorithm for Data Compression.


Minimum Error Rate Training in Statistical Machine Translation.

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

Modelling and Optimizing on Syntactic N-Grams for Statistical Machine Translation.

How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs.


Edinburgh Neural Machine Translation Systems for WMT 16.

Improving Neural Machine Translation Models with Monolingual Data.

Neural Machine Translation of Rare Words with Subword Units.

A tree does not make a well-formed sentence: Improving syntactic string-to-tree statistical machine translation with more linguistic knowledge.

Hybrid Machine Translation: integration of linguistics and statistics.

Attention Is All You Need.
CoRR, abs/1706.03762.