Machine Translation with Diverse Data Sources

Huda Khayrallah

This talk was presented at JHU CLSP seminar on March 29, 2019 and at the UPenn Computational Linguistics Seminar on April 8, 2019.

It is based on the following papers:

https://aclweb.org/anthology/W18-2705
(bibtex: https://aclweb.org/anthology/W18-2705)

https://aclweb.org/anthology/W18-2709
(bibtex: https://aclweb.org/anthology/W18-2709.bib)
Machine Translation with Diverse Data Sources

Huda Khayrallah

Work with:
Brian Thompson, Kevin Duh & Philipp Koehn
Overview

• Review of Neural Machine Translation (NMT)
• Review of Domain Adaptation
• Improving Domain Adaptation
  • Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation [Khayrallah, Thompson, Duh & Koehn 2018]
• Analysis of Noisy Corpora
  • On the Impact of Various Types of Noise on Neural Machine Translation [Khayrallah & Koehn 2018]
Neural Machine Translation
Wasch dir die Hände
Wasch dir die Hände
Wasch dir die Hände

Encoder

Source Embedding

Wasch dir die Hände
Wasch dir die Hände
Wasch dir die Hände
Wash  

Softmax  

Decoder  

Encoder  

Source Embedding  

Wasch dir die Hände
Wash

Target Embedding

Softmax

Decoder

Encoder

Source Embedding

Source: Wasch, dir, die, Hände
Wash

Target Embedding

Softmax

Decoder

Encoder

Source Embedding

Wasch  dir  die  Hände
Wash
your hands

Target Embedding

Softmax

Decoder

Encoder

Source Embedding

Wasch dir die Hände

Wash your hands
NMT loss function

\[ \mathcal{L}_{\text{NLL}}(\theta) = - \sum_{v \in \mathcal{V}} \left( \mathbb{1}\{y_i = v\} \times \log p(y_i = v \mid x; \theta; y_j < i) \right) \]

Cross Entropy (Gold Target, Model output)
Overview

• Review of Neural Machine Translation (NMT)

• **Review of Domain Adaptation**

• Improving Domain Adaptation
  • Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation [Khayrallah, Thompson, Duh & Koehn 2018]

• Analysis of Noisy Corpora
  • On the Impact of Various Types of Noise on Neural Machine Translation [Khayrallah & Koehn 2018]
What do we want to translate?
Developmental toxicity, including dose-dependent delayed foetal ossification and possible teratogenic effects, were observed in rats at doses resulting in subtherapeutic exposures (based on AUC) and in rabbits at doses resulting in exposures 3 and 11 times the mean steady-state AUC at the maximum recommended clinical dose.
The films coated therewith, in particular polycarbonate films coated therewith, have improved properties with regard to scratch resistance, solvent resistance, and reduced oiling effect, said films thus being especially suitable for use in producing plastic parts in film insert molding methods.
General Domain Data
General Domain Data

Would it not be beneficial, in the short term, following the Rotterdam model, to inspect according to a points system in which, for example, account is taken of the ship's age, whether it is single or double-hulled or whether it sails under a flag of convenience.
Mama always said there's an awful lot you can tell about a person by their shoes.
Domain Mismatch
Translating Russian Patents
General Domain NMT

50m General Domain sentence pairs

General Domain NMT Model
дверной замок повышенной степени защищенности от взлома
Human: door lock with increased degree of security against burglary
System: door security door security door
In-Domain NMT

30k In-Domain sentence pairs

In-Domain NMT Model
In-Domain NMT

Errors due to lack of data

dверной замок повышенной степени защищенности от взлома
Human: door lock with increased degree of security against burglary
System: door lock for a high degree of protection against coke
Domain Adaptation
Continued Training

50m General Domain sentence pairs

General Domain NMT Model

30k In-domain sentence pairs

Continued Training NMT Model
Continued Training

дверной замок повышенной степени защищенности от взлома
Human: door lock with increased degree of security against burglary
System: door lock with increased penetration protection

Improved performance!

General Domain NMT Model

30k In-domain sentence pairs

Continued Training NMT Model

Huda Khayrallah
Russian → English Patents

General Domain

In-Domain

Continued Training + 9.3

BLEU
Overview

• Review of Neural Machine Translation (NMT)
• Review of Domain Adaptation
• Improving Domain Adaptation
  • Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation [Khayrallah, Thompson, Duh & Koehn 2018]
• Analysis of Noisy Corpora
  • On the Impact of Various Types of Noise on Neural Machine Translation [Khayrallah & Koehn 2018]
Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation

Huda Khayrallah, Brian Thompson, Kevin Duh & Philipp Koehn
WNMT at ACL 2018
Continued Training

General Domain NMT Model

30k In-domain sentence pairs

Continued Training NMT Model

Huda Khayrallah
Regularized Continued Training

General Domain NMT Model → 30k In-domain sentence pairs → Regularized Continued Training NMT Model

Huda Khayrallah
Teacher/Student Models

• Word Level Knowledge distillation

• Often used to make smaller/faster models
• Train one model; use it to ‘teach’ another
Regularized Continued Training

Teacher

General Domain NMT Model

Student

Regularized Continued Training NMT Model

General Domain NMT Model
NMT loss function

\[
\mathcal{L}_{\text{NLL}}(\theta) = - \sum_{v \in \mathcal{V}} (\mathbb{1}\{y_i = v\} \times \log p(y_i = v | x; \theta; y_j < i))
\]

Cross Entropy (Gold Target, CT Model output)
Teacher/Student Loss Function

\[- \sum_{v \in \mathcal{V}} (p_{aux}(y_i = v \mid x; \theta_{aux}; y_j < i) \times \log(p(y_i = v \mid x; \theta; y_j < i)) \]

General Model Output (teacher)

Cross Entropy (,)

General Model Output (teacher)

CT Model output (student)

CT Model output (student)
This work: Combine Both

\[(1 - \alpha) \times \left( - \sum_{v \in V} \left( \mathbb{1}\{y_i = v\} \times \log p(y_i = v \mid x; \theta; y_{j<i}\) \right) \right) + \]

\[\alpha \times \left( - \sum_{v \in V} \left( p_{aux}(y_i = v \mid x; \theta_{aux}; y_{j<i}) \times \log p(y_i = v \mid x; \theta; y_{j<i}) \right) \right)\]

\[(1 - \alpha) \times \text{Cross Ent } (\text{Orange Set}, \text{Purple Set}) + \]

\[\alpha \times \text{Cross Ent } (\text{Blue Set}, \text{Purple Set})\]
Results
Russian → English Patents

![Graph showing BLEU scores for different domains and training methods.]

- General Domain: Blue bar
- In-Domain: Red bar
- Continued Training: Light blue bar
- Continued Training w/Reg: Purple bar

BLEU scores:
- General Domain: 20
- In-Domain: 30
- Continued Training: 25
- Continued Training w/Reg: 31.2

Increase: +1.2
English → German Medical

<table>
<thead>
<tr>
<th>Domain</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Domain</td>
<td></td>
</tr>
<tr>
<td>In-Domain</td>
<td>40</td>
</tr>
<tr>
<td>Continued Training w/Reg</td>
<td>+1.5</td>
</tr>
</tbody>
</table>

Huda Khayrallah
Analysis
Russian → English General (patents)

![Bar chart showing BLEU scores for General Domain and Continued Training with and without regularization.](chart.png)

- General Domain: -9.2
- Continued Training: -18.2
- Continued Training w/Reg: -9.2
Overcoming Catastrophic Forgetting During Domain Adaptation of Neural Machine Translation

Brian Thompson†  Jeremy Gwinnup°  Huda Khayrallah†  Kevin Duh†  Philipp Koehn†

†Johns Hopkins University, °Air Force Research Laboratory
{brian.thompson, huda, phi}@jhu.edu,
kevinduh@cs.jhu.edu,
jeremy.gwinnup.1@us.af.mil
Overview

• Review of Neural Machine Translation (NMT)
• Review of Domain Adaptation
• Improving Domain Adaptation
  • Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation [Khayrallah, Thompson, Duh & Koehn 2018]

• Analysis of Noisy Corpora
  • On the Impact of Various Types of Noise on Neural Machine Translation [Khayrallah & Koehn 2018]
On the Impact of Various Types of Noise on Neural Machine Translation

Huda Khayrallah & Philipp Koehn
WNMT at ACL 2018 [Outstanding Contribution Award]
More data is better!

Figure 3: BLEU scores for English-Spanish systems.

Figure 4: Translations of the first sentence of the test set using NMT system trained on varying amounts of training data.

---

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

- **Statistical MT (SMT)**
- **Statistical MT with big LM**
- **Neural MT (NMT)**

---

Corpus size (English words)
Let’s go get more data!
# De⇒En translation

<table>
<thead>
<tr>
<th></th>
<th>NMT</th>
<th>SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT17</td>
<td>27.2</td>
<td>24.0</td>
</tr>
<tr>
<td>+ raw paracrawl</td>
<td>17.3 (-9.9)</td>
<td>25.2 (+1.2)</td>
</tr>
</tbody>
</table>
Raw Paracrawl

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>27.4</td>
<td>26.6</td>
<td>24.7</td>
<td>20.9</td>
<td>17.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+0.2</td>
<td>+0.6</td>
<td>+2.5</td>
<td>+6.3</td>
<td>-6.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>24.2</td>
<td>24.2</td>
<td>24.4</td>
<td>24.8</td>
<td>25.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+0.2</td>
<td>+0.2</td>
<td>+0.4</td>
<td>+0.8</td>
<td>+1.2</td>
<td></td>
</tr>
</tbody>
</table>

**NMT**

**SMT**
Manual Analysis

Okay

Misaligned sentences

Other Text

Short Segments

3rd Language

Untranslated

Both German

Both English

Huda Khayrallah
Noise Types

• Misaligned Sentences
• Misordered words
• Wrong Language
• Untranslated Sentences
• Short Segments
Misaligned Sentences
Misaligned Sentences

Die Koalas sind süß  The koalas are cute
Die Kängurus springen  The kangaroos jump
Der Koala ist weich  The koala is soft
Das Känguru ist schnell  The kangaroo is fast
Die Koalas sind süß
Die Kängurus springen
Der Koala ist weich
Das Känguru ist schnell

The kangaroos jump
The koala is soft
The kangaroo is fast
The koalas are cute
## Misaligned Sentences

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>50%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOURCE</td>
<td>26.5</td>
<td>24.0</td>
<td>26.3</td>
<td>26.1</td>
<td>25.3</td>
</tr>
<tr>
<td></td>
<td>-0.7</td>
<td>-0.0</td>
<td>-0.9</td>
<td>-1.1</td>
<td>-1.9</td>
</tr>
<tr>
<td>TARGET</td>
<td>26.5</td>
<td>24.0</td>
<td>23.9</td>
<td>23.9</td>
<td>23.4</td>
</tr>
<tr>
<td></td>
<td>-0.7</td>
<td>-0.0</td>
<td>-0.1</td>
<td>-0.1</td>
<td>-0.6</td>
</tr>
</tbody>
</table>

**NMT**  **SMT**
Misordered Words
Misordered Words (source)

Die Koalas sind süß
The koalas are cute

Die Kängurus springen
The kangaroos jump

Der Koala ist weich
The koala is soft

Das Känguru ist schnell
The kangaroo is fast
Misordered Words (source)

Koalas Die sind süß  The koalas are cute
Kängurus springen Die  The kangaroos jump
ist Der weich Koala  The koala is soft
schnell Känguru ist Das  The kangaroo is fast
## Misordered Words (source)

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>50%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26.9</td>
<td>24.0</td>
<td>26.6</td>
<td>23.6</td>
<td>26.6</td>
<td>23.6</td>
</tr>
<tr>
<td>-0.3</td>
<td>-0.0</td>
<td>-0.6</td>
<td>-0.4</td>
<td>-0.6</td>
<td>-0.4</td>
</tr>
<tr>
<td>26.4</td>
<td>23.9</td>
<td>26.6</td>
<td>23.6</td>
<td>25.5</td>
<td>23.7</td>
</tr>
<tr>
<td>-0.8</td>
<td>-0.1</td>
<td>-0.6</td>
<td>-0.4</td>
<td>-1.7</td>
<td>-0.3</td>
</tr>
</tbody>
</table>
Misordered Words (target)

Die Koalas sind süß 
Die Kängurus springen
Der Koala ist weich
Das Känguru ist schnell

The koalas are cute
The kangaroos jump
The koala is soft
The kangaroo is fast
Misordered Words (target)

Die Koalas sind süß  koalas cute are The
Die Kängurus springen  kangaroos The jump
Der Koala ist weich  is The soft koala
Das Känguru ist schnell  fast The is kangaroo
## Misordered Words (target)

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>50%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>5%</strong></td>
<td>27.0</td>
<td>26.8</td>
<td>26.4</td>
<td>26.7</td>
<td>26.1</td>
</tr>
<tr>
<td></td>
<td>-0.2</td>
<td>-0.4</td>
<td>-0.8</td>
<td>-0.5</td>
<td>-1.1</td>
</tr>
<tr>
<td><strong>10%</strong></td>
<td>24.0</td>
<td>24.0</td>
<td>23.4</td>
<td>23.2</td>
<td>22.9</td>
</tr>
<tr>
<td></td>
<td>-0.0</td>
<td>-0.0</td>
<td>-0.6</td>
<td>-0.8</td>
<td>-1.1</td>
</tr>
<tr>
<td><strong>20%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>50%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>100%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**NMT**  **SMT**
Wrong Language
Wrong Language (French source)

Die Koalas sind süß  The koalas are cute
Die Kängurus springen  The kangaroos jump
Der Koala ist weich  The koala is soft
Das Känguru ist schnell  The kangaroo is fast
Les koalas sont mignons  The koalas are cute
Les kangourous sautent  The kangaroos jump
Le koala est doux  The koala is soft
Le kangourou est rapide  The kangaroo is fast
Wrong Language (French source)

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>50%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>26.9</td>
<td>26.8</td>
<td>26.8</td>
<td>26.8</td>
<td>26.8</td>
</tr>
<tr>
<td></td>
<td>24.0</td>
<td>23.9</td>
<td>23.9</td>
<td>23.9</td>
<td>23.8</td>
</tr>
<tr>
<td></td>
<td>-0.3</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.4</td>
</tr>
<tr>
<td></td>
<td>-0.0</td>
<td>-0.1</td>
<td>-0.1</td>
<td>-0.1</td>
<td>-0.2</td>
</tr>
</tbody>
</table>
Wrong Language (French target)

Die Koalas sind süß
The koalas are cute

Die Kängurus springen
The kangaroos jump

Der Koala ist weich
The koala is soft

Das Känguru ist schnell
The kangaroo is fast
Wrong Language (French target)

Die Koalas sind süß  Les koalas sont mignons
Die Kängurus springen  Les kangourous sautent
Der Koala ist weich  Le koala est doux
Das Känguru ist schnell  Le kangourou est rapide
Wrong Language (French target)

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>50%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>26.7</td>
<td>24.0</td>
<td>26.6</td>
<td>23.9</td>
<td>26.7</td>
<td>23.8</td>
</tr>
<tr>
<td>-0.5</td>
<td>-0.0</td>
<td>-0.6</td>
<td>-0.1</td>
<td>-0.5</td>
<td>-0.2</td>
</tr>
</tbody>
</table>

NMT  SMT
Untranslated
Untranslated (English Source)

Die Koalas sind süß
The koalas are cute

Die Kängurus springen
The kangaroos jump

Der Koala ist weich
The koala is soft

Das Känguru ist schnell
The kangaroo is fast
Untranslated (English source)

The koalas are cute
The kangaroos jump
The koala is soft
The kangaroo is fast
## Untranslated (English source)

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>50%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>5%</strong></td>
<td>27.2 23.9</td>
<td>27.0 23.9</td>
<td>26.7 23.6</td>
<td>26.8 23.7</td>
<td>26.9 23.5</td>
</tr>
<tr>
<td><strong>10%</strong></td>
<td>-0.0 -0.1</td>
<td>-0.2 -0.1</td>
<td>-0.5 -0.4</td>
<td>-0.4 -0.3</td>
<td>-0.3 -0.5</td>
</tr>
<tr>
<td><strong>20%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>50%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>100%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NMT  SMT
Die Koalas sind süß
Die Kängurus springen
Der Koala ist weich
Das Känguru ist schnell

The koalas are cute
The kangaroos jump
The koala is soft
The kangaroo is fast
Die Koalas sind süß
Die Kängurus springen
Der Koala ist weich
Das Känguru ist schnell

Die Koalas sind süß
Die Kängurus springen
Der Koala ist weich
Das Känguru ist schnell
### Untranslated (German target)

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>50%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>5%</strong></td>
<td>17.6</td>
<td>11.2</td>
<td>5.6</td>
<td>3.2</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>23.8</td>
<td>23.9</td>
<td>23.8</td>
<td>23.4</td>
<td>21.1</td>
</tr>
<tr>
<td></td>
<td>-0.2</td>
<td>-0.1</td>
<td>-0.2</td>
<td>-0.6</td>
<td>-2.9</td>
</tr>
<tr>
<td><strong>10%</strong></td>
<td>-9.8</td>
<td>-16.0</td>
<td>-21.6</td>
<td>-24.0</td>
<td>-24.0</td>
</tr>
</tbody>
</table>

- **NMT**
- **SMT**
Short Segments
Short Segments

Die süß Känguru schnell

The cute Kangaroo fast
# Short Segments

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 2 words</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>27.1</td>
<td>24.1</td>
<td>26.5</td>
<td>23.9</td>
</tr>
<tr>
<td></td>
<td>-0.1</td>
<td>+0.1</td>
<td>-0.7</td>
<td>-0.1</td>
</tr>
<tr>
<td>3-5 words</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>27.8</td>
<td>24.2</td>
<td>27.6</td>
<td>24.5</td>
</tr>
<tr>
<td></td>
<td>+0.6</td>
<td>+0.2</td>
<td>+0.4</td>
<td>+0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 1: Results from adding different amounts of noise (ratio of original clean corpus) for various types of noise in German-English Translation. Generally neural machine translation (left green bars) is harmed more than statistical machine translation (right blue bars). The worst type of noise are segments in the source language copied untranslated into the target.

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>50%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Misaligned Sentences</td>
<td>26.5 24.0</td>
<td>26.5 24.0</td>
<td>26.3 23.9</td>
<td>26.1 23.9</td>
<td>25.3 23.4</td>
</tr>
<tr>
<td></td>
<td>-0.7 -0.0</td>
<td>-0.7 -0.0</td>
<td>-0.9 -0.1</td>
<td>-1.1 -0.1</td>
<td>-1.9 -0.6</td>
</tr>
<tr>
<td>Misordered Words (Source)</td>
<td>26.9 24.0</td>
<td>26.6 23.6</td>
<td>26.4 23.9</td>
<td>26.6 23.6</td>
<td>25.5 23.7</td>
</tr>
<tr>
<td></td>
<td>-0.3 -0.0</td>
<td>-0.6 -0.4</td>
<td>-0.8 -0.1</td>
<td>-0.6 -0.4</td>
<td>-1.7 -0.3</td>
</tr>
<tr>
<td>Misordered Words (Target)</td>
<td>27.0 24.0</td>
<td>26.8 24.0</td>
<td>26.4 23.4</td>
<td>26.7 23.2</td>
<td>26.1 22.9</td>
</tr>
<tr>
<td></td>
<td>-0.2 -0.0</td>
<td>-0.4 -0.0</td>
<td>-0.8 -0.6</td>
<td>-0.5 -0.8</td>
<td>-1.1 -1.1</td>
</tr>
<tr>
<td>Wrong Language (French Source)</td>
<td>26.9 24.0</td>
<td>26.8 23.9</td>
<td>26.8 23.9</td>
<td>26.8 23.9</td>
<td>26.8 23.8</td>
</tr>
<tr>
<td></td>
<td>-0.3 -0.0</td>
<td>-0.4 -0.1</td>
<td>-0.4 -0.1</td>
<td>-0.4 -0.1</td>
<td>-0.4 -0.2</td>
</tr>
<tr>
<td>Wrong Language (French Target)</td>
<td>26.7 24.0</td>
<td>26.6 23.9</td>
<td>26.7 23.8</td>
<td>26.2 23.5</td>
<td>25.0 23.4</td>
</tr>
<tr>
<td></td>
<td>-0.5 -0.0</td>
<td>-0.6 -0.1</td>
<td>-0.5 -0.2</td>
<td>-1.0 -0.5</td>
<td>-2.2 -0.6</td>
</tr>
<tr>
<td>Untranslated (English Source)</td>
<td>27.2 23.9</td>
<td>27.0 23.9</td>
<td>26.7 23.6</td>
<td>26.8 23.7</td>
<td>26.9 23.5</td>
</tr>
<tr>
<td></td>
<td>-0.0 -0.1</td>
<td>-0.2 -0.1</td>
<td>-0.5 -0.4</td>
<td>-0.4 -0.3</td>
<td>-0.3 -0.5</td>
</tr>
<tr>
<td>Untranslated (German Target)</td>
<td>17.6 23.8</td>
<td>11.2 23.9</td>
<td>5.6 23.8</td>
<td>3.2 23.4</td>
<td>3.2 21.1</td>
</tr>
<tr>
<td></td>
<td>-0.2 -16.0</td>
<td>-0.1 -21.6</td>
<td>-0.2 -24.0</td>
<td>-0.6 -24.0</td>
<td>-2.9 -24.0</td>
</tr>
<tr>
<td>Short Segments (max 2)</td>
<td>27.1 24.1</td>
<td>26.5 23.9</td>
<td>26.7 23.8</td>
<td>26.7 23.8</td>
<td>26.7 23.8</td>
</tr>
<tr>
<td></td>
<td>-0.1 +0.1</td>
<td>-0.7 -0.1</td>
<td>-0.5 -0.2</td>
<td>-0.5 -0.2</td>
<td>-0.5 -0.2</td>
</tr>
<tr>
<td>Short Segments (max 5)</td>
<td>27.8 24.2</td>
<td>27.6 24.5</td>
<td>28.0 24.5</td>
<td>26.6 24.2</td>
<td>26.6 24.2</td>
</tr>
<tr>
<td></td>
<td>+0.6 +0.4</td>
<td>+0.4 +0.5</td>
<td>+0.8 +0.5</td>
<td>-0.6 +0.2</td>
<td>-0.6 +0.2</td>
</tr>
<tr>
<td>Raw Crawl Data</td>
<td>27.4 24.2</td>
<td>26.6 24.2</td>
<td>24.7 24.4</td>
<td>20.9 24.8</td>
<td>17.3 25.2</td>
</tr>
<tr>
<td></td>
<td>+0.2 +0.2</td>
<td>-0.6 +0.2</td>
<td>-2.5 +0.4</td>
<td>+0.4 -6.3</td>
<td>+1.2 -9.9</td>
</tr>
</tbody>
</table>
Filtering methods

- BiCleaner [Espla-Gomis & Forcada 2009]
- Zipporah [Xu & Koehn 2017]
- WMT shared task [Koehn, Khayrallah, Heafield & Forcada 2018]
  - Dual Conditional Cross-Entropy Filtering [Junczys-Dowmunt 2018]
  - Zipporah [Khayrallah, Xu & Koehn 2018]
Overview

• Review of Neural Machine Translation (NMT)
• Review of Domain Adaptation
• Improving Domain Adaptation
  • Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation [Khayrallah, Thompson, Duh & Koehn 2018]
• Analysis of Noisy Corpora
  • On the Impact of Various Types of Noise on Neural Machine Translation [Khayrallah & Koehn 2018]
Questions?