Machine Translation with Diverse Data Sources

Huda Khayrallah

This talk was presented at NYU Abu Dhabi CS seminar on September 1, 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
huda@jhu.edu

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

• Overview of Neural Machine Translation (NMT)
• Overview 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]
How do I get to the airport?

كيف يمكنني الوصول إلى المطار؟

kayf yumkinuni alwusul 'iilaa almatar?
Neural Machine Translation
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
Wasch dir die Hände
Wasch dir die Hände

Source Embedding

Wesch dir die Hände
Wasch dir die Hände
Wasch dir die Hände

Decoder

Encoder

Source Embedding

Wasch, dir, die, Hände
Wasch dir die Hände

Softmax
Decoder
Encoder
Source Embedding

Wasch dir die Hände
Softmax
Decoder
Encoder
Source Embedding

Wasch  dir  die  Hände
Wasch dir die Hände
Wash

Target Embedding

Softmax

Decoder

Encoder

Source Embedding

Wasch dir die Hände
Wash

Target Embedding

Softmax

Decoder

Encoder

Source Embedding

Wasch dir die Hände
Wash your hands.
NMT loss function

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

Cross Entropy( Gold Target, Model output )

Gold Target

Model output
BLEU

• Weighted n-grams precision

\[
\min \left( 1, \frac{\text{output length}}{\text{reference length}} \right) \left( \prod_{i=1}^{4} \text{precision}_i \right)^{\frac{1}{4}}
\]

• Between 0 and 1
  • (often scaled to be 0-100)

• Higher is better

• Imperfect...

• But... not bad

Khayrallah
Overview

• Overview of Neural Machine Translation (NMT)

• **Overview 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
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
Case Study:
Translating Russian Patents
General Domain NMT

50m General Domain sentence pairs

General Domain NMT Model

Huda Khayrallah
Human: door lock with increased degree of security against burglary

System: door security door security door

Errors due to domain mismatch
In-Domain NMT

30k In-Domain sentence pairs

In-Domain NMT Model
In-Domain NMT

30k In-Domain sentence pairs

In-Domain NMT Model

Errors due to lack of data

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

Improved performance!

50m General Domain sentence pairs

дверной замок повышенной степени защищенности от взлома

Human: door lock with increased degree of security against burglary

System: door lock with increased penetration protection
Russian → English Patents

General Domain

In-Domain

Continued Training

+ 9.3

BLEU

General Domain

In-Domain

Continued Training

Johns Hopkins University
Russian → English General

<table>
<thead>
<tr>
<th>BLEU</th>
<th>General Domain</th>
<th>Continued Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>-18.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Overview

• Overview of Neural Machine Translation (NMT)
• Overview 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

General Domain NMT Model
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

General Domain NMT Model

30k In-domain sentence pairs

Regularized Continued Training NMT Model

Student

Teacher

General Domain NMT Model
NMT loss function

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

Cross Entropy(\quad, \quad )

Gold Target \quad CT Model output

Gold Target \quad CT Model output
Teacher/Student Loss Function

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

Cross Entropy (General Model Output (teacher), CT Model output (student))
This work: Combine Both

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

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

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

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

BLEU

General Domain
Continued Training
In-Domain
Continued Training w/Reg

+ 1.2
German → English Medical

<table>
<thead>
<tr>
<th>BLEU</th>
<th>General Domain</th>
<th>Continued Training</th>
<th>In-Domain</th>
<th>Continued Training w/Reg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>+ 0.8</td>
<td></td>
</tr>
</tbody>
</table>

Huda Khayrallah
English → German Medical

BLEU

+ 1.5

General Domain
Continued Training
In-Domain
Continued Training w/Reg

Huda Khayrallah
English → German Medical

BLEU

+ 1.5

General Domain
Continued Training

In-Domain
Continued Training w/Reg

Huda Khayrallah
Russian → English General (patents)

![Bar chart showing BLEU scores for different training methods: General Domain, Continued Training, and Continued Training w/ Reg. General Domain has a BLEU score of -18.2, Continued Training has a BLEU score of -9.2.](image-url)
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
German-English Medical – Small

BLEU

+ 0.2

General Domain

Continued Training

Continued Training w/Reg
English-German Medical – Small

General Domain

Continued Training

Continued Training w/Reg

+ 0.2
Overview

• Overview of Neural Machine Translation (NMT)
• Overview 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]
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>
</tbody>
</table>

+ noisy corpus

WMT17: 27.2 ± 9.9
SMT: 25.2 ± 1.2
More data is better!

Figure 3: BLEU scores for English-Spanish systems.

Figure 4: Translations of the first sentence of Edinburgh's WMT submission

To illustrate this, see Figure 4, which shows translations of the first sentence of Edinburgh's WMT submission.

The contrast between the NMT and SMT learning curves is quite striking. While NMT is able to exploit increasing amounts of training data more effectively, it is unable to get off the ground with much lower, outperforming SMT at about 15 million words. For SMT, the training corpus sizes of a few million words or less are highly-inflected categories. Conventional wisdom states that neural machine translation models perform particularly poorly on rare words, but do continue to have difficulty translating some.

Figure 4

Some key words are properly translated when there is sufficient training data, the output is completely unrelated to the input, some key words are properly translated.

To examine this claim by comparing performance on smaller vocabularies used by NMT systems. We find that NMT systems actually outperform SMT systems on translation of very infrequent words. However, both NMT and SMT systems use a publicly available model trained on all additional provided monolingual data and SMT system trained on each subset, we also use all additionally provided monolingual data.

BLEU for SMT, 30.4 for SMT+BigLM. A model with the full data set (31.1 for NMT, 28.4 even beating the SMT system with a big language model) with Spanish. To obtain a learning curve, we used a publicly available model trained on the Spanish part of the data)

...abysmal results (BLEU score of 1.6 vs. 16.4 for NMT systems trained on 0.4 million to 385.7 million words of parallel data. Quality for NMT starts much lower, outperforms SMT at about 385.7 million English words paired (24.1 million words), outperforming SMT 25.7 vs. 24.7 with bleeding word similarity in embeddings) and comparing? NMT promises both to generalize better (except under high-resource conditions.

Figure 4: Translations of the first sentence of Edinburgh’s WMT submission (Koehn & Knowles 2017)
More data is better!

![Graph showing BLEU scores vs. corpus size (English words)](image)

**Figure 3:** BLEU scores for English-Spanish systems.

**Figure 4:** Translations of the first sentence of Arthur et al. (2016a).
More data is better!

![Graph showing BLEU score increase with corpus size](graph.png)

**Table 2:** German-English IWSLT results for training corpus size of 100k words and 3.2M words (full corpus). Mean and standard deviation of three training runs reported.

**Figure 2:** German-English learning curve, showing BLEU as a function of the amount of parallel training data, for PBSMT and NMT.
Let’s go get more data!
Annual growth in prices came in at 10.9 per cent, more than double the gain of the 12 months to August 2013, but the gains were not evenly spread across the country.
### 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</td>
<td>25.2</td>
</tr>
</tbody>
</table>

- (red) indicates a decrease compared to WMT17.
- (green) indicates an increase compared to WMT17.
## Raw Paracrawl

<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>5%</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.4</td>
<td>24.2</td>
<td>26.6</td>
<td>24.7</td>
<td>20.9</td>
<td>17.3</td>
</tr>
<tr>
<td>+0.2</td>
<td>+0.2</td>
<td>-0.6</td>
<td>+0.4</td>
<td>+0.8</td>
<td>+1.2</td>
</tr>
</tbody>
</table>

NMT  SMT
Manual Analysis

- Okay
- Misaligned sentences
- Other Text
- Short Segments
- 3rd Language
- Untranslated
- Both German
- Both English
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
Misaligned Sentences

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

<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>5%</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Misaligned Sentences</td>
<td>26.5</td>
<td>24.0</td>
<td>26.5</td>
<td>24.0</td>
<td>26.3</td>
</tr>
<tr>
<td>Misordered Words</td>
<td>26.9</td>
<td>24.0</td>
<td>26.6</td>
<td>23.6</td>
<td>26.4</td>
</tr>
<tr>
<td>Wrong Language (French Source)</td>
<td>26.9</td>
<td>24.0</td>
<td>26.8</td>
<td>23.9</td>
<td>26.8</td>
</tr>
<tr>
<td>Wrong Language (French Target)</td>
<td>26.7</td>
<td>24.0</td>
<td>26.6</td>
<td>23.9</td>
<td>26.4</td>
</tr>
<tr>
<td>Untranslated (English Source)</td>
<td>27.2</td>
<td>23.9</td>
<td>27.0</td>
<td>23.9</td>
<td>26.7</td>
</tr>
<tr>
<td>Untranslated (German Target)</td>
<td>17.6</td>
<td>23.8</td>
<td>11.2</td>
<td>23.9</td>
<td>5.6</td>
</tr>
<tr>
<td>Short Segments (max 2)</td>
<td>27.1</td>
<td>24.1</td>
<td>26.5</td>
<td>23.9</td>
<td>26.7</td>
</tr>
<tr>
<td>Short Segments (max 5)</td>
<td>27.8</td>
<td>24.2</td>
<td>27.6</td>
<td>24.5</td>
<td>28.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NMT</th>
<th>SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5%</td>
<td>-0.7</td>
</tr>
<tr>
<td>5%</td>
<td>-0.7</td>
</tr>
<tr>
<td>10%</td>
<td>-0.9</td>
</tr>
<tr>
<td>25%</td>
<td>-1.1</td>
</tr>
<tr>
<td>50%</td>
<td>-1.9</td>
</tr>
</tbody>
</table>
Misordered Words
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 (source)

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

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

<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>5%</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>26.9</td>
<td>24.0</td>
<td>26.6</td>
<td>23.6</td>
<td>25.5</td>
</tr>
<tr>
<td></td>
<td>-0.3</td>
<td>-0.0</td>
<td>-0.6</td>
<td>-0.4</td>
<td>-1.7</td>
</tr>
<tr>
<td>Target</td>
<td>26.4</td>
<td>23.9</td>
<td>26.6</td>
<td>23.6</td>
<td>23.7</td>
</tr>
<tr>
<td></td>
<td>-0.8</td>
<td>-0.1</td>
<td>-0.6</td>
<td>-0.4</td>
<td>-0.3</td>
</tr>
<tr>
<td>English</td>
<td>26.1</td>
<td>23.9</td>
<td>25.3</td>
<td>23.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-1.1</td>
<td>-0.6</td>
<td>-1.9</td>
<td>-0.6</td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>26.8</td>
<td>24.0</td>
<td>26.8</td>
<td>24.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.4</td>
<td>-0.0</td>
<td>-0.4</td>
<td>-0.0</td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>26.6</td>
<td>23.9</td>
<td>26.8</td>
<td>23.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.8</td>
<td>-0.1</td>
<td>-0.8</td>
<td>-0.1</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>26.2</td>
<td>23.5</td>
<td>25.0</td>
<td>23.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-1.0</td>
<td>-0.8</td>
<td>-2.2</td>
<td>-1.1</td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>26.7</td>
<td>23.8</td>
<td>26.8</td>
<td>23.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.5</td>
<td>-0.2</td>
<td>-0.4</td>
<td>-0.2</td>
<td></td>
</tr>
</tbody>
</table>

**NMT**  **SMT**
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 (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>2.5%</th>
<th>5%</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.0</td>
<td>-0.2</td>
<td>26.8</td>
<td>26.4</td>
<td>26.7</td>
<td>26.1</td>
</tr>
<tr>
<td>24.0</td>
<td>-0.0</td>
<td>24.0</td>
<td>23.4</td>
<td>23.2</td>
<td>22.9</td>
</tr>
<tr>
<td>-0.2</td>
<td>-0.4</td>
<td>-0.8</td>
<td>-0.5</td>
<td>-1.1</td>
<td></td>
</tr>
<tr>
<td>-0.0</td>
<td>-0.0</td>
<td>-0.6</td>
<td>-0.8</td>
<td>-1.1</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
Wrong Language (French source)

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>2.5%</th>
<th>5%</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT</td>
<td>26.9</td>
<td>26.8</td>
<td>26.8</td>
<td>26.8</td>
<td>26.8</td>
</tr>
<tr>
<td>SMT</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
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
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>2.5%</th>
<th>5%</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>26.7</td>
<td>26.6</td>
<td>26.7</td>
<td>26.2</td>
<td>25.0</td>
</tr>
<tr>
<td>24.0</td>
<td>23.9</td>
<td>23.8</td>
<td>23.5</td>
<td>23.4</td>
</tr>
<tr>
<td>-0.5</td>
<td>-0.6</td>
<td>-0.5</td>
<td>-1.0</td>
<td>-2.2</td>
</tr>
<tr>
<td>-0.0</td>
<td>-0.1</td>
<td>-0.2</td>
<td>-0.5</td>
<td>-0.6</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
The koalas are cute
The kangaroos jump
The koala is soft
The kangaroo is fast
Untranslated (English source)

<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>5%</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>27.2</td>
<td>27.0</td>
<td>26.7</td>
<td>26.8</td>
<td>26.9</td>
</tr>
<tr>
<td></td>
<td>23.9</td>
<td>23.9</td>
<td>23.6</td>
<td>23.7</td>
<td>23.5</td>
</tr>
<tr>
<td></td>
<td>-0.0</td>
<td>-0.2</td>
<td>-0.5</td>
<td>-0.4</td>
<td>-0.3</td>
</tr>
<tr>
<td></td>
<td>-0.1</td>
<td>-0.1</td>
<td>-0.4</td>
<td>-0.3</td>
<td>-0.5</td>
</tr>
</tbody>
</table>

NMT    SMT
Untranslated (German 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
Untranslated (German target)

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>2.5%</th>
<th>5%</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT Source</td>
<td>17.6</td>
<td>11.2</td>
<td>5.6</td>
<td>3.2</td>
<td>3.2</td>
</tr>
<tr>
<td>NMT Target</td>
<td>23.8</td>
<td>23.9</td>
<td>23.8</td>
<td>23.4</td>
<td>21.1</td>
</tr>
<tr>
<td>SMT Source</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMT Target</td>
<td>-0.2</td>
<td>-0.1</td>
<td>-0.2</td>
<td>-0.6</td>
<td>-2.9</td>
</tr>
</tbody>
</table>

Huda Khayrallah
Short Segments
Short Segments

Die süß Känguru schnell
The cute Kangaroo fast
# Short Segments

<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>5%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>≤ 2 words</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>27.1</td>
<td>24.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.1</td>
<td>+0.1</td>
<td></td>
</tr>
<tr>
<td><strong>3-5 words</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>27.8</td>
<td>24.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+0.6</td>
<td>+0.2</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>5%</td>
<td>10%</td>
<td>20%</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>Misaligned Sentences</td>
<td>26.5 24.0</td>
<td>26.5 24.0</td>
<td>26.3 23.9</td>
</tr>
<tr>
<td></td>
<td>-0.7 -0.0</td>
<td>-0.7 -0.0</td>
<td>-0.9 -0.1</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>
</tr>
<tr>
<td></td>
<td>-0.3 -0.0</td>
<td>-0.6 -0.4</td>
<td>-0.8 -0.1</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>
</tr>
<tr>
<td></td>
<td>-0.2 -0.0</td>
<td>-0.4 -0.0</td>
<td>-0.8 -0.6</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>
</tr>
<tr>
<td></td>
<td>-0.3 -0.0</td>
<td>-0.4 -0.1</td>
<td>-0.4 -0.1</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>
</tr>
<tr>
<td></td>
<td>-0.5 -0.0</td>
<td>-0.6 -0.1</td>
<td>-0.5 -0.2</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>
</tr>
<tr>
<td></td>
<td>-0.0 -0.1</td>
<td>-0.2 -0.1</td>
<td>-0.5 -0.4</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>
</tr>
<tr>
<td></td>
<td>-9.8 -16.0 -21.6 -24.0 -24.0</td>
<td></td>
<td></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>
</tr>
<tr>
<td></td>
<td>-0.1 +0.1</td>
<td>-0.7 -0.1</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>
</tr>
<tr>
<td></td>
<td>+0.6 +0.2</td>
<td>+0.4 +0.5</td>
<td>+0.8 +0.5</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>
</tr>
<tr>
<td></td>
<td>+0.2 +0.2</td>
<td>-0.6 +0.2</td>
<td>-2.5 +0.4</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.
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]
<table>
<thead>
<tr>
<th></th>
<th>NMT</th>
<th>SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WMT17</strong></td>
<td>27.2</td>
<td>24.0</td>
</tr>
<tr>
<td><strong>+ raw paracrawl</strong></td>
<td>17.3((-9.9))</td>
<td>25.2((+1.2))</td>
</tr>
</tbody>
</table>
### 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>
<tr>
<td>WMT19 + filtered paracrawl</td>
<td>32.4 ((+5.2))</td>
<td>25.8 ((+1.8))</td>
</tr>
</tbody>
</table>
Overview

• Overview of Neural Machine Translation (NMT)
• Overview 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]
Takeaways

• new methods $\rightarrow$ improved performance
• models consist of methods + **data**
• methods pick up patterns in the data
  • noise [Khayrallah & Koehn 2018]
  • annotation artifacts [Poliak et al., 2018]
  • Bias [Ethics in NLP 2017, 2018; Gender Bias in NLP 2019]
Questions?

huda@jhu.edu

cs.jhu.edu/~huda