Neural Machine Translation
what’s linguistics got to do with it?

Rico Sennrich
University of Edinburgh
Setting the Scene: 2014–2015

research trend: more linguistics for statistical machine translation

Laura hat einen kleinen Garten
Laura has a small garden

syntax-based LM
[Sennrich, TACL 2015]

morphological structure
[Sennrich, Haddow, EMNLP 2015]

a new challenger appears: neural machine translation

- requires minimal domain knowledge
- similar models used for speech and computer vision
Edinburgh’s* WMT Results over the Years

![Graph showing Edinburgh's WMT Results over the years from 2013 to 2015.

- **2013**:
  - Phrase-based SMT: 20.3
  - Syntax-based SMT: 19.4

- **2014**:
  - Phrase-based SMT: 20.9
  - Syntax-based SMT: 20.2

- **2015**:
  - Phrase-based SMT: 20.8
  - Syntax-based SMT: 22.0

*NMT 2015 from U. Montréal: [https://sites.google.com/site/acl16nmt/](https://sites.google.com/site/acl16nmt/)

Rico Sennrich | NMT: what’s linguistics got to do with it? 2/39
Edinburgh’s* WMT Results over the Years

*NMT 2015 from U. Montréal: https://sites.google.com/site/acl16nmt/

Rico Sennrich

NMT: what’s linguistics got to do with it?

2/39
Edinburgh’s* WMT Results over the Years

![Bar chart showing BLEU scores for different years and MT methods.](chart.png)

*BLEU scores for Edinburgh's WMT Results over the years (2013-2017).

**Methods:**
- Phrase-based SMT
- Syntax-based SMT
- Neural MT

*NMT 2015 from U. Montréal: [https://sites.google.com/site/acl16nmt/](https://sites.google.com/site/acl16nmt/)

Rico Sennrich  |  NMT: what’s linguistics got to do with it?  |  2/39
What Now?

do we still need linguistics for MT?
What Now?

do we still need linguistics for MT?
What Now?

do we still need linguistics for MT?
Today’s Talk

areas in which linguistics is helping neural MT research

- linguistically motivated (but non-linguistic) models
- linguistically informed models
- targeted evaluation of neural MT
source reference
indoor temperature Raumklima
[Bahdanau et al., 2015] UNK ✓
[Jean et al., 2015] Innenpool ✓
[Sennrich, Haddow, Birch, ACL 2016a] Innen+ temperatur ✓

Figure 4: Example development sentence, showing the inferred attention matrix for various models for Et ↔ En. Rows correspond to the translation direction and columns correspond to different models: attentional, with alignment features (+align), global fertility (+glofer), and symmetric joint training (+sym). Darker shades denote higher values and white denotes zero.

Pre). In this case, it is refining an already excellent model from which reliable global fertility estimates can be obtained. This finding is consistent with the other languages, see Figure 3 which shows typical learning curves of different variants of the attentional model. Note that when global fertility is added to the vanilla attentional model with alignment features, it significantly slows down training as it limits exploration in early training iterations, however it does bring a sizeable win when used to fine-tune a pre-trained model. Finally, the bilingual symmetry also helps to reduce the perplexity scores when used with the alignment features, however, does not combine well with global fertility (+align+sym+glofer-pre). This is perhaps an unsurprising result as both methods impose a often-times similar regularising effect over the attention matrix.

Figure 4 illustrates the different attention matrices inferred by the various model variants. Note the difference between the base attentional model and its variant with alignment features (+align), where more weight is assigned to diagonal and 1-to-many alignments. Global fertility pushes more attention to the sentinel symbols 〈s〉 and 〈/s〉. Determiners and prepositions in English show much lower fertility than nouns, while Estonian nouns have even higher fertility. This accords with Estonian morphology, wherein nouns are inflected with rich case marking, e.g., nõukoguga has the cogitative -ga suffix, meaning 'with', and thus translates as several English words (with the council).

The rightmost column corresponds to joint symmetric training, with many more confident attention values especially for consistent 1-to-many alignments (difficult in English and raskeid in Estonian, an adjective in partitive case meaning some difficult).
Linguistic Structure is Coming Back to (Neural) MT

<table>
<thead>
<tr>
<th>Segmentation</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>perusasian</td>
</tr>
<tr>
<td>BPE</td>
<td>perusasi: an</td>
</tr>
<tr>
<td>Omorfi</td>
<td>perus: asia: n</td>
</tr>
</tbody>
</table>

Morphology

[Sánchez-Cartagena and Toral, 2016]
[Tamchyna et al., 2017]
[Huck et al., 2017]
[Pinnis et al., 2017]

Syntax

[Sennrich and Haddow, 2016]
[Eriguchi et al., 2016]
[Bastings et al., 2017]
[Aharoni and Goldberg, 2017]
[Nadejde et al., 2017]
Targeted Evaluation of Neural MT
hypothesis: model A obtains higher BLEU than model B on data set X
What Hypotheses Do We Test?

hypothesis: model A obtains higher BLEU than model B on data set X
What Hypotheses Do We Test?

hypothesis: model A is better model of translation than model B

evidence: model A obtains higher BLEU than model B on data set X
What Hypotheses Do We Test?

hypothesis: model A is better model of translation than model B

evidence: model A obtains higher BLEU than model B on data set X
| hypothesis: | many languages have long-distance interactions. model A produces disfluent output because it models these interactions poorly. model B can better model long-distance interactions, and produces more fluent output. |
What Hypotheses Do We Test?

<table>
<thead>
<tr>
<th>hypothesis:</th>
</tr>
</thead>
<tbody>
<tr>
<td>many languages have long-distance interactions.</td>
</tr>
<tr>
<td>model A produces disfluent output because it models these interactions poorly.</td>
</tr>
<tr>
<td>model B can better model long-distance interactions, and produces more fluent output.</td>
</tr>
</tbody>
</table>

Francis Victoria Gumapac / CC BY 2.0
What Hypotheses Do We Test?

| hypothesis: | many languages have long-distance interactions. model A produces disfluent output because it models these interactions poorly. model B can better model long-distance interactions, and produces more fluent output. |
| evidence:   | model A obtains higher BLEU than model B on data set X |
What Hypotheses Do We Test?

| hypothesis: | many languages have long-distance interactions. model A produces disfluent output because it models these interactions poorly. model B can better model long-distance interactions, and produces more fluent output. |
| evidence:   | model A obtains higher BLEU than model B on data set X |
being able to test our hypotheses is beauty of empirical NLP

- complex, interesting hypotheses need targeted evaluation
- I want to see more interesting hypotheses
  → we need more targeted evaluation
Fluency
is translation good English?
+13%

Adequacy
is meaning preserved?
+1%

Figure: WMT16 direct assessment results
Human Evaluation in TraMOOC

- direct assessment of NMT (vs. PBSMT):
  - fluency: +10%
  - adequacy: +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>
Human Evaluation of Neural MT

Neural Machine Translation is very fluent.
Human Evaluation of Neural MT

Neural Machine Translation is very fluent.

Attentional encoder-decoder with BPE segmentation and recurrent GRU decoder
Human Evaluation of Neural MT

**Neural Machine Translation** is very fluent.

Attentional encoder-decoder with BPE segmentation and recurrent GRU decoder

**what about...?**

- character-level models [Lee et al., 2016]
- convolutional models [Gehring et al., 2017]
- models with self-attention [Vaswani et al., 2017]
Neural Machine Translation is very fluent.

Attentional encoder-decoder with BPE segmentation and recurrent GRU decoder

what about...?

- character-level models [Lee et al., 2016]
- convolutional models [Gehring et al., 2017]
- models with self-attention [Vaswani et al., 2017]

Adequacy remains a major problem in Neural Machine Translation
Human Evaluation of Neural MT

Neural Machine Translation is very fluent.

Attentional encoder-decoder with BPE segmentation and recurrent GRU decoder

what about...?

- character-level models [Lee et al., 2016]
- convolutional models [Gehring et al., 2017]
- models with self-attention [Vaswani et al., 2017]

Adequacy remains a major problem in Neural Machine Translation

...using a shallow NMT model at WMT 2016
Human Evaluation of Neural MT

Neural Machine Translation is very fluent.

Attentional encoder-decoder with BPE segmentation and recurrent GRU decoder

what about...?

- character-level models [Lee et al., 2016]
- convolutional models [Gehring et al., 2017]
- models with self-attention [Vaswani et al., 2017]

Adequacy remains a major problem in Neural Machine Translation

...using a shallow NMT model at WMT 2016

how...?

- do we compare different architectures?
- do we measure improvement over time?
How to Assess Specific Aspects in MT?

- human evaluation
  - costly; hard to compare to previous work
- automatic metrics (BLEU)
  - too coarse; blind towards specific aspects
How to Assess Specific Aspects in MT?

- human evaluation
  × costly; hard to compare to previous work
- automatic metrics (BLEU)
  × too coarse; blind towards specific aspects

**contrastive translation pairs**

- NMT models assign probability to any translation
- binary classification task: which translation is better?
- choice between reference translation and contrastive variant
  → corrupted with single error of specific type
- ≈ minimal pairs in linguistics
Assessment with Contrastive Translation Pairs

workflow

- researcher wants to analyse difficult translation problem
- researcher predicts what errors NMT system might make
- researcher creates test set with correct translations and corrupted variants
- test set allows automatic, quantitative, and reproducible analysis of NMT model

example
Assessment with Contrastive Translation Pairs

**Workflow**
- researcher wants to analyse difficult translation problem
- researcher predicts what errors NMT system might make
- researcher creates test set with correct translations and corrupted variants
- test set allows automatic, quantitative, and reproducible analysis of NMT model

**Example**
- subject–verb agreement

---

Rico Sennrich  
NMT: what’s linguistics got to do with it?  
16/39
Assessment with Contrastive Translation Pairs

**workflow**
- researcher wants to analyse difficult translation problem
- researcher predicts what errors NMT system might make
- researcher creates test set with correct translations and corrupted variants
- test set allows automatic, quantitative, and reproducible analysis of NMT model

**example**
- subject–verb agreement
- change grammatical number of verb to introduce agreement error
Assessment with Contrastive Translation Pairs

**Workflow**
- researcher wants to analyse difficult translation problem
- researcher predicts what errors NMT system might make
- researcher creates test set with correct translations and corrupted variants
- test set allows automatic, quantitative, and reproducible analysis of NMT model

**Example**
- subject–verb agreement
- change grammatical number of verb to introduce agreement error
- 35000 contrastive pairs created with simple linguistic rules
<table>
<thead>
<tr>
<th>Language</th>
<th>Sentence</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>[...] that the <strong>plan will</strong> be approved</td>
<td>0.1</td>
</tr>
<tr>
<td>German (correct)</td>
<td>[...] , dass der <strong>Plan</strong> verabschiedet <strong>wird</strong></td>
<td></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*
LingEval97: A Test Set of Contrastive Translation Pairs

- 97,000 contrastive translation pairs
- Based on English $\rightarrow$ German WMT test sets
- Rule-based, automatic creation of errors
- 7 error types
- Metadata for in-depth analysis:
  - Error type
  - Distance between words
  - Word frequency in WMT15 training set
Case Study: Some Open Questions in Neural MT

Kyunghyun Cho
@kchonyc

Fully char-level NMT! It works well on all four language pairs we've considered ({Cs, De, Ru, Fi}→En), and we... fb.me/1oRwyQvZD
Case Study: Some Open Questions in Neural MT

Fully char-level NMT! It works well on all four language pairs we've considered (Cs, De, Ru, Fi)->En), and we... fb.me/1oRwyQvZD

@kchonyc Are there any benefits to using these models for longer dependencies?
Case Study: Some Open Questions in Neural MT

Fully char-level NMT! It works well on all four language pairs we’ve considered (Cs, De, Ru, Fi)->En, and we... fb.me/1oRwyQvZD

Emiel van Miltenburg
@evanmiltenburg

@kchonyc Are there any benefits to using these models for longer dependencies?

Kyunghyun Cho
@kchonyc

@evanmiltenburg ah well that’s a difficult question!
<table>
<thead>
<tr>
<th>Level</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word-level</td>
<td>but as the <strong>example</strong> of MobilKing in Poland <strong>shows</strong></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Subword-level (byte-pair encoding)</td>
<td>but as the <strong>example</strong> of Mobil+ king in Poland <strong>shows</strong></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Character-level</td>
<td>but as the <strong>example</strong> of MobilKing in Poland <strong>shows</strong></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Case Study: Some Open Questions in Neural MT

Fully char-level NMT! It works well on all four language pairs we've considered ({Cs, De, Ru, Fi}→En), and we... fb.me/1oRwyQvZD

|————– 5 steps —————|

subword-level (byte-pair encoding)

but as the example of Mobil+ king in Poland shows
|——— 6 steps ————|

character-level

but as the example of UNK in Poland shows
|——— 29 steps ————|

@kchonyc Are there any benefits to using these models for longer dependencies?
1:16 PM - 11 Oct 2016

@evanmiltenburg ah well that's a difficult question!
1:30 PM - 11 Oct 2016

Rico Sennrich
NMT: what's linguistics got to do with it? 19/39
Case Study: Some Open Questions in Neural MT

does network architecture affect learning of long-distance dependencies?

architectures

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.

RNN vs. GRU vs. LSTM

Christopher Olah http://colah.github.io/posts/2015-08-Understanding-LSTMs/
Case Study: Some Open Questions in Neural MT

does network architecture affect learning of long-distance dependencies?

architectures

RNN vs. GRU vs. LSTM
(1) convolution

[Gebrung et al., 2017]

(2) self-attention

[Vaswani et al., 2017]

---

Christopher Olah
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Rico Sennrich
NMT: what’s linguistics got to do with it?
Results: Architecture

subject-verb agreement
n=35105

accuracy (%)

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN</td>
<td>89.0</td>
</tr>
<tr>
<td>GRU</td>
<td>93.4</td>
</tr>
<tr>
<td>LSTM</td>
<td>94.0</td>
</tr>
</tbody>
</table>

Rico Sennrich

NMT: what’s linguistics got to do with it?
Results: Architecture

Accuracy (subject-verb agreement) vs. distance:

- LSTM
- GRU
- RNN

Accuracy decreases with increasing distance.
Results: Text Representation

Subject-verb agreement
n=35105

Accuracy (%)

- BPE-to-BPE: 93.4%
- Word-to-word: 90.5%
- Char-to-char [Lee et al., 2016]: 91.5%

Rico Sennrich
NMT: what’s linguistics got to do with it? 22/39
Results: Text Representation

![Graph showing accuracy vs. frequency](image)

- **BPE-to-BPE**
- **word-to-word**
- **char-to-char [Lee et al., 2016]**

Accuracy (subject-verb agreement) vs. frequency.
Results: Text Representation

![Graph showing accuracy (subject-verb agreement) vs. distance]

- **BPE-to-BPE**
- **word-to-word**
- **char-to-char [Lee et al., 2016]**

The graph illustrates the accuracy of subject-verb agreement for different distances, comparing BPE-to-BPE, word-to-word, and char-to-char approaches. The accuracy remains relatively high for distances up to 16, with slight variations observed.
What Did We Learn?

- method verifies strength of LSTM and GRU
  - future work: test of convolutional model and self-attention
- word-level model is poor for rare words
- character-level model is poor for long distances
- BPE subword segmentation is good compromise
### adequacy is open problem

<table>
<thead>
<tr>
<th>system</th>
<th>sentence</th>
</tr>
</thead>
</table>
| source       | 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. |
| reference    |                                                                           |
| our NMT      | There he was attacked again by the **racket** and another male person.    |
| Google       | There he was again attacked by the **bat** and another male person.       |

**Schläger**
Targeted Analysis: Adequacy

adequacy is open problem

<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. There he was attacked again by his original attacker and another male.</td>
</tr>
<tr>
<td>reference</td>
<td>Our NMT</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

attacker
adequacy is open problem

<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>
### Adequacy is an open problem

<table>
<thead>
<tr>
<th>System</th>
<th>Source Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>system</td>
<td>sentence</td>
</tr>
<tr>
<td>source</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>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**
- **racket**
- **attacker**
- **bat**
### Targeted Analysis: Adequacy

<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>source reference</td>
<td>There he was attacked again by the <strong>racket</strong> and another male person.</td>
</tr>
<tr>
<td>our NMT Google</td>
<td>There he was again attacked by the <strong>bat</strong> and another male person.</td>
</tr>
</tbody>
</table>

#### Diagram

- **Schläger**
- **racket**
- **attacker**
- **bat**
Targeted Analysis: Adequacy

focus on two types of adequacy errors:

- lexical word sense disambiguation:
  translate ambiguous word with wrong word sense

- polarity:
  deletion or insertion of negation marker ("not", "no", "un-")
manual error analysis [Fancellu and Webber, 2015]

translation errors (Chinese $\rightarrow$ English hierarchical PBSMT):

- insertion of negation (1–2%)
- deletion of negation (10–20%)
- reordering errors (1–20%)
manual error analysis [Fancellu and Webber, 2015]

translation errors (Chinese → English hierarchical PBSMT):
- insertion of negation (1–2%)
- deletion of negation (10–20%)
- reordering errors (1–20%)

automatic analysis (Lingeval97; NMT)

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation insertion</td>
<td>97.9</td>
</tr>
<tr>
<td>Negation deletion</td>
<td>91.5</td>
</tr>
</tbody>
</table>

n=22760, n=4043
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

<table>
<thead>
<tr>
<th>source:</th>
<th>Also nahm ich meinen amerikanischen Reisepass und stellte mich in die Schlange für Extranjeros.</th>
</tr>
</thead>
<tbody>
<tr>
<td>reference:</td>
<td>So I took my U.S. passport and got in the <strong>line</strong> for Extranjeros.</td>
</tr>
<tr>
<td>contrastive:</td>
<td>So I took my U.S. passport and got in the <strong>snake</strong> for Extranjeros.</td>
</tr>
<tr>
<td>contrastive:</td>
<td>So I took my U.S. passport and got in the <strong>serpent</strong> for Extranjeros.</td>
</tr>
</tbody>
</table>
Word Sense Accuracy

WSD is challenging, especially for rare word senses.
Word Sense Accuracy

WSD is challenging, especially for rare word senses
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

Word sense disambiguation accuracy
n=7359

Accuracy (%)

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

![Graph showing word sense disambiguation accuracy](image)

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

- accuracy (%)
- word sense disambiguation accuracy
- n=7359

Accuracy:
- UEDIN-NMT @ WMT16: 83.5%
- UEDIN-NMT @ WMT17: 86.7%
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, 83.5%
- UEDIN-NMT @ WMT17: single, 86.7%
- UEDIN-NMT @ WMT17: ensemble, 87.9%
- ≈ human performance (sentence-level), 96.0%

≈ human performance (sentence-level)
What Did We Learn?

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

  German reference  
  Sehen Sie die *Muster*?
  contrastive reference  
  Do you see the *patterns*?

  → new possibility for targeted evaluation of document-level modelling
### background

Antecedent agreement can often not be predicted based on source sentence, but requires extra-sentential context:

<table>
<thead>
<tr>
<th>English</th>
<th>I made a decision.</th>
<th>Please respect it.</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>J’ai pris une décision.</td>
<td>Respectez-la s’il vous plaît.</td>
</tr>
<tr>
<td>French</td>
<td>J’ai fait un choix.</td>
<td>Respectez-le s’il vous plaît.</td>
</tr>
</tbody>
</table>
Targeted Analysis: Coreference

Previous work: shared task on pronoun prediction

[Hardmeier et al., 2015, Guillou et al., 2016, Loáiciga et al., 2017]

- Focus on correctness of pronouns, which are often coreferent.
- Pronoun errors impact meaning, but only have small effect on BLEU.
- Limitations of shared task:
  - Many pronouns do not require extra-sentential context; sentence-level system still best at DiscoMT17 [Loáiciga et al., 2017].
  - We want to analyze NMT systems’ ability to model coreference, without training specifically for this task, but:
    - Task gives lemmatized target side
    - Long tail of possible pronouns handled via OTHER category
Contrastive Pairs for Analysis of Coreference in MT

[Bawden, Sennrich, Birch, Haddow, in preparation]

Source:
context: Oh, I hate flies. Look, there's another one!
current sent.: Don’t worry, I'll kill it for you.

Target:
1 context: Ô je déteste les mouches. Regarde, il y en a une autre !
correct: T'inquiète, je la tuerai pour toi.
incorrect: T'inquiète, je le tuerai pour toi.

2 context: Ô je déteste les moucherons. Regarde, il y en a un autre !
correct: T'inquiète, je le tuerai pour toi.
incorrect: T'inquiète, je la tuerai pour toi.

design of test set
- hand-crafted set of 200 contrastive pairs
- previous sentence required for correct prediction
- balanced so that sentence-level system scores 50%
## Coreference Models

### baseline setup
- training on OpenSubtitles EN-FR [Tiedemann, 2012]
- attentional encoder-decoder (Nematus) with BPE

### architectures
- sentence-level baseline
- 2-TO-1: concatenation of previous source sentence
- 2-TO-2: concatenation of previous source and target sentence
- S-MULTI: separate encoder for previous source; hierarchical attention
- S-MULTI-TO-2: separate encoder for previous source; previous target sentence concatenated

### related work
- [Tiedemann and Scherrer, 2017] (2-TO-*)
- [Zoph and Knight, 2016, Libovický and Helcl, 2017] (S-MULTI)
Targeted Analysis: Coreference: Results

Pronoun prediction accuracy

n=200

50.0 %

Accuracy (%)

pronoun prediction accuracy

n=200

baseline
2-to-1
2-to-2
s-multi
s-multi-to-2

Rico Sennrich

NMT: what’s linguistics got to do with it?
Targeted Analysis: Coreference: Results

Pronoun prediction accuracy
n=200

50.0 52.0 63.5

Accuracy (%)

baseline 2-to-1 2-to-2 s-multi s-multi-to-2

Rico Sennrich  NMT: what’s linguistics got to do with it?
Targeted Analysis: Coreference: Results

pronoun prediction accuracy
n=200

accuracy (%)
40.0 50.0 60.0 70.0 80.0 90.0 100.0

baseline 2-to-1 2-to-2 s-multi 2-to-2 s-multi-to-2

Rico Sennrich NMT: what’s linguistics got to do with it?
## Coreference Models: BLEU Results

<table>
<thead>
<tr>
<th>System</th>
<th>Comedy</th>
<th>Crime</th>
<th>Fantasy</th>
<th>Horror</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single-encoder, non-contexual model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BASELINE</td>
<td>19.52</td>
<td>22.07</td>
<td>26.30</td>
<td>33.05</td>
</tr>
<tr>
<td><strong>Single-encoder with concatenated input</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-TO-2</td>
<td>20.09</td>
<td><strong>22.93</strong></td>
<td>26.60</td>
<td>33.59</td>
</tr>
<tr>
<td>2-TO-1</td>
<td>19.51</td>
<td>21.81</td>
<td>26.78</td>
<td><strong>34.37</strong></td>
</tr>
<tr>
<td><strong>Multi-encoder, multi-attention models (+previous source sentence)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-MULTI</td>
<td><strong>20.22</strong></td>
<td>21.90</td>
<td>26.81</td>
<td>34.04</td>
</tr>
<tr>
<td><strong>Multi-encoder, multi-attention models with concatenated output</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-MULTI-TO-2</td>
<td><strong>20.85</strong></td>
<td>22.81</td>
<td>27.17</td>
<td><strong>34.62</strong></td>
</tr>
</tbody>
</table>
What Did We Learn?

- target context is crucial for prediction of correct pronoun (partially due to test set, in which source words are ambiguous)
- targeted evaluation can guide our exploration of architectures → multi-encoder architecture only works in some conditions (*-to-2)
neural machine translation does not need linguistic knowledge...
...but linguistics should play an important role for inspiring research

informing models

targeted evaluation

source reference
[Bahdanau et al., 2015]
Jean et al., 2015
[Sennrich, Haddow, Birch, ACL 2016a]
indoor temperature Raumklima
UNK
Innenpool
Innen+ temperatur

racket
attacker
bat

Schläger

h(i)

h(j)

h(k)

*PAD*

Thou
monkey
cats
tennis

importance of edges, separate weight matrices are used as in a non-directional GCN.

For an alternative approach to integrating labels and directionality, we arrive at:

\[
\hat{w}_{uv} = \sigma(\sum_{j} \rho((W^j)_{dir(u,v)} h_j(u) + b_j)_{lab(u,v)}))
\]

where \( \hat{w}_{uv} \) is the logistic sigmoid function, and \( \rho \) is the exponential function.

Longer dependencies can be modeled with edge-wise gates which let the model regulate contributions of individual dependency edges. We use linguistic structures such as dependency trees, where directionality and edge labels play an important role. They also integrate as dependency trees, where directionality and edge labels play an important role for incoming and outgoing edges. We follow the convention that in dependency trees heads point to their dependents, and thus outgoing edges are used for dependent-to-head connections, and incoming edges are used for head-to-dependent connections, and thus outgoing edges are used for dependent-to-head connections, and incoming edges are used for head-to-dependent connections. Making the GCN sensitive to labels is straightforward given the above modifications for edges.

This allows us to operate on directed and labeled graphs. Gates and some labels are omitted for clarity.

In order to deal with directionality, we arrive at:

\[
\hat{w}_{uv} = \sigma(\sum_{j} \rho((W^j)_{dir(u,v)} h_j(u) + b_j)_{lab(u,v)}))
\]

In order to deal with directionality, we arrive at:

Figure 2: A 2-layer syntactic GCN on top of a convolutional encoder. Loop connections are depicted with dashed edges, syntactic ones with solid (dependents to heads) and dotted (heads to dependents) edges. Gates and some labels are omitted for clarity.
Collaborators

Joint work with:

Alexandra Birch
Annette Rios

Barry Haddow
Laura Mascarell

Rachel Bawden
Acknowledgments

Some of the research presented was conducted in cooperation with Samsung Electronics Polska.

This work has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreements 645452 (QT21), TraMOOC (644333), HimL (644402), and SUMMA (688139).

This work has received funding from the Swiss National Science Foundation (SNF) in the project CoNTra (grant number 105212_169888).
Thank you for your attention

Resources

- LingEval97: https://github.com/rsennrich/lingeval97
- ContraWSD: https://github.com/a-rios/ContraWSD
- Discourse test set: https://diamt.limsi.fr/eval.html
- Pre-trained models:
  - WMT16: http://data.statmt.org/wmt16_systems/
  - WMT17: http://data.statmt.org/wmt17_systems/
Towards String-To-Tree Neural Machine Translation.

Neural Machine Translation by Jointly Learning to Align and Translate.

Graph Convolutional Encoders for Syntax-aware Neural Machine Translation.
Proceedings of EMNLP.


Glyph-aware Embedding of Chinese Characters.
In 1st Workshop on Subword and Character level models in NLP (SCLeM), Copenhagen, Denmark.

A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators.
In Proceedings of Machine Translation Summit XVI, Nagoya, Japan.
Incorporating Structural Alignment Biases into an Attentional Neural Translation Model.

Chinese–Spanish neural machine translation enhanced with character and word bitmap fonts.

Tree-to-Sequence Attentional Neural Machine Translation.


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

Findings of the 2016 WMT Shared Task on Cross-lingual Pronoun Prediction.


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

The University of Edinburgh’s Neural MT Systems for WMT17.


Linguistic Input Features Improve Neural Machine Translation.

Edinburgh Neural Machine Translation Systems for WMT 16.

Neural Machine Translation of Rare Words with Subword Units.
Modeling Target-Side Inflection in Neural Machine Translation.
In Second Conference on Machine Translation (WMT17).

Tiedemann, J. (2012).
Parallel Data, Tools and Interfaces in OPUS.

Neural Machine Translation with Extended Context.

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


Multi-Source Neural Translation.
In NAACL HLT 2016.