Teaching Machine Translation
additional Constraints

Jan Niehues

23/06/2020
jan.niehues@maastrichtuniversity.nl
Motivation

• NMT reach very good quality
  • Condition
    • Large amount of training data
    • Similar domain of training and test data

• Real-world applications
  - Often additional constrained necessary
    - Length constraints
    - Time constraints
  - No training data available
Length-constrained machine translation

- Generate translation with a given length
  - Focus on shortening

- Translations of websites
  - Fit into layout

- Subtitles
  - Cognitive load
    - Adjust to reading speed
Time-constrained machine translation

- Live transcription
  - Cannot wait for full sentence

- Strategies to output intermediate outputs
  - Update previous outputs
  - Dynamically decide when to output

- Latency:
  - Time between spoken words and display of the translation
Overview

• Motivation

• Length Constraints

• Readability in subtitles

• Low-latency sequence-to-sequence models
Length-constrained translation

- **Aim:**
  - User is able to control length of translation

- **Input:**
  - Source language sentence
  - Desired target length

- **Output:**
  - Target sentence fulfilling length contained
    - Soft/Hard constraints

- **Variants**
  - Mono-lingual translation/Paraphrasing
Baseline

- Restrict search space
  - Only generate hypothesis fulfilling length constraint
  - Limit has research, increase probability of \(<\text{s}>\)

- Hard constraint
- No modification of training

- Problems:
  - Beginning of sentence cannot be changed
Ideas

- Length aware during the whole generation
  - Plan your available spots
  - Shorten already at the beginning

- Challenges:
  - Target length also known during training
    - Training data with length
  - How to integrate length into model
Pseudo-supervised training

- **Goal:**
  - Training data with given target length
  - Available with different length ratios

- **Challenge:**
  - Hard to acquire

<table>
<thead>
<tr>
<th>Source</th>
<th>Es klingt vielleicht übel.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>8</td>
</tr>
<tr>
<td>Reference</td>
<td>4</td>
</tr>
<tr>
<td>Reference</td>
<td>4</td>
</tr>
</tbody>
</table>
Pseudo-supervised training

- **Idea:**
  - Assume parallel training data was generated using length constraints

- **Advantage:**
  - Hugh amounts of training data for different domains

- **Disadvantage:**
  - Model might ignore length information

<table>
<thead>
<tr>
<th>Source</th>
<th>Es klingt vielleicht übel.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>It might sound like it's a bad thing.</td>
</tr>
</tbody>
</table>
Length representation

• Use length as additional input to the encoder
  • Successfully done in multi-lingual MT, domain adaptation,…

• Challenge:
  - Long lengths might be rare (e.g. 63)
  - Model might ignore length due to long distance from loss
Length representation

• Integrate into decoder
  • More direct influence on output probability

• Use remaining length at each step of the decoding process
  • Countdown to sentence end
  • Similar to positional encoding
Length representation

- Include length information into the initial representation of each target work

- Embedding
  - Concatenate embedding for the remaining length
Length representation

- Include length information into the initial representation of each target work
- Embedding
  - Concatenate embedding for the remaining length
- Positional encoding
  - Encode remaining length instead of position
Evaluation

• No available evaluation data
• Use automatic metrics against original reference
• Problem:
  - Word-based metrics
  - Embedding-based metric
    - RUSE

<table>
<thead>
<tr>
<th>Reference</th>
<th>It might sound like it's a bad thing.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>But it might sound like</td>
</tr>
<tr>
<td>Constraint</td>
<td>It sounds really bad</td>
</tr>
</tbody>
</table>
System

• IWSLT Multi-lingual set (2017)
  - German, English, Italian, Dutch and Romanian
  - Or German-English subset
  - Standard preprocessing with BPE
  - Target length: 80% and 50% of the source sentence

• Transformer
  - 8-layers
  - 512/2048
Task difficulty

- Force reference length

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>RUSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>30.80</td>
<td>-0.085</td>
</tr>
<tr>
<td>Only Search</td>
<td>28.32</td>
<td>-0.124</td>
</tr>
<tr>
<td>Source Emb</td>
<td>28.56</td>
<td>-0.126</td>
</tr>
<tr>
<td>Decoder Emb</td>
<td>27.88</td>
<td>-0.140</td>
</tr>
<tr>
<td>Decoder Pos</td>
<td>28.80</td>
<td>-0.138</td>
</tr>
</tbody>
</table>
# Length representation

- RUSE scores

<table>
<thead>
<tr>
<th>Model</th>
<th>80%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>-0.272</td>
<td>-0.605</td>
</tr>
<tr>
<td>Source Emb</td>
<td>-0.263</td>
<td>-0.587</td>
</tr>
<tr>
<td>Decoder Emb</td>
<td>-0.247</td>
<td>-0.555</td>
</tr>
<tr>
<td>Decoder Pos</td>
<td>-0.260</td>
<td>-0.577</td>
</tr>
</tbody>
</table>

Jan Niehues - Teaching Machine Translation additional Constraints
Multi-lingual

- English-English as zero-shot translation of multi-lingual machine translation system
  - Target Length 80%

<table>
<thead>
<tr>
<th>Model</th>
<th>Baseline</th>
<th>Decoder Emb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE-EN</td>
<td>-0.225</td>
<td>-0.214</td>
</tr>
<tr>
<td>EN-EN</td>
<td>-0.102</td>
<td>0.020</td>
</tr>
</tbody>
</table>
Cascade vs. End-to-End

- Target length 80%

<table>
<thead>
<tr>
<th>Model</th>
<th>DE-EN</th>
<th>EN-EN</th>
</tr>
</thead>
<tbody>
<tr>
<td>End2-End</td>
<td>-0.247</td>
<td>0.020</td>
</tr>
<tr>
<td>Cascade</td>
<td>-0.259</td>
<td>-0.118</td>
</tr>
<tr>
<td>Cascade Fix. Pivot</td>
<td>-0.259</td>
<td>-0.118</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.166</td>
</tr>
</tbody>
</table>
Examples

<table>
<thead>
<tr>
<th>Source</th>
<th>Und, obwohl es wirklich einfach scheint, ist es tatsächlich richtig schwer, weil es Leute drängt sehr schnell zusammenzuarbeiten.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>And, though it seems really simple, it's actually pretty hard because it forces people to collaborate very quickly.</td>
</tr>
<tr>
<td>Base 0.8</td>
<td>and even though it really seems simple, it is actually really hard, because it really pushes</td>
</tr>
<tr>
<td>Dec. Emb 0.8</td>
<td>and although it really seems simple, it is really hard because it drives people to work together.</td>
</tr>
<tr>
<td>Base 0.5</td>
<td>and even though it really seems simple, it is really hard</td>
</tr>
<tr>
<td>Dec. Emb 0.5</td>
<td>it is really hard because it drives people to work together.</td>
</tr>
</tbody>
</table>
Simplification

- Can we use the same framework for other tasks?
- Simplification:
  - Assumptions:
    - Words split by BPE are complex
    - Minimize split words
  - Approach:
    - Only count sub words
    - Generate translation with target length 0
## Simplification - Result

<table>
<thead>
<tr>
<th>Metric</th>
<th>Base</th>
<th>Simplified</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPE tokens</td>
<td>1899</td>
<td>991</td>
</tr>
<tr>
<td>DCI</td>
<td>7.66</td>
<td>7.45</td>
</tr>
<tr>
<td>BLEU</td>
<td>32.84</td>
<td>31.29</td>
</tr>
</tbody>
</table>
Readability

- Until now:
  - Compared to default translation

- Comparison to human subtitles
  - Generate for German TV News

- Monolingual
  - Aligned with audio
"Befreit vom fraktionszwang soll das Parlament wohl nach der Sommerpause die ethisch schwierige Frage debattieren."
summer break the ethically difficult question debate."
"Without party-constraints should the parliament maybe after the summer break the ethically difficult question debate."
Experiments

• Casaded
  - First ASR
  - Then compression

• End-to-End
  - Transcription & compression in one model
### Datasets

- **Unsupervised compression model**
  - \{de, en, it, nl, ro\} TED talks from IWSLT 2017

- **End-to-end model**

<table>
<thead>
<tr>
<th>Partitions</th>
<th>Total length (h:m)</th>
<th>Total utterances</th>
</tr>
</thead>
<tbody>
<tr>
<td>LibriVoxDeEn (train)</td>
<td>469:21</td>
<td>206,490</td>
</tr>
<tr>
<td>Tagesshau (adapt)</td>
<td>37:28</td>
<td>11,559</td>
</tr>
<tr>
<td>Tagesshau (test)</td>
<td>46</td>
<td>213</td>
</tr>
</tbody>
</table>

Beilharz et al. (2020). LibriVoxDeEn: A Corpus for German-to-English Speech Translation and German Speech Recognition. Proc. LREC

Jan Niehues - Teaching Machine Translation additional Constraints
Results

![Bar chart showing results for WER, R-1, R-2, and R-L with off-the-shelf ASR and +Compress]

- WER ↓
- R-1 ↑
- R-2 ↑
- R-L ↑

Off-the-shelf ASR +Compress
Results

**Spoken:** Es ist kurz nach Mitternacht, als plötzlich ein Auto in eine Gruppe von Menschen steuert, die ausgelassen ins neue Jahr feiern.

It is shortly after midnight, when suddenly a car into a group of people drives, that happily into new year celebrate.

**Ref:** kurz nach Mitternacht steuert ein Auto in eine Gruppe von Menschen, die ins neue Jahr feiern.

**Output:** kurz nach Mitternacht fährt ein Auto plötzlich in eine Gruppe von Leuten, die das nächste Jahr feiern.
Results

Spoken: Es ist kurz nach Mitternacht, als plötzlich ein Auto in eine Gruppe von Menschen steuert, die ausgelassen ins neue Jahr feiern.

It is shortly after midnight, when suddenly a car into a group of people drives, that happily into new year celebrate.

Ref: kurz nach Mitternacht steuert ein Auto in eine Gruppe von Menschen, die ins neue Jahr feiern.

Output: kurz nach Mitternacht fährt ein Auto plötzlich in eine Gruppe von Leuten, die das nächste Jahr feiern.
**Results**

**Spoken:** Es ist kurz nach Mitternacht, als plötzlich ein Auto in eine Gruppe von Menschen steuert, die ausgelassen ins neue Jahr feiern.

It is shortly after midnight, when suddenly a car into a group of people drives, that happily into new year celebrate.

**Ref:** kurz nach Mitternacht steuert ein Auto in eine Gruppe von Menschen, die ins neue Jahr feiern.

**Output:** kurz nach Mitternacht fährt ein Auto plötzlich in eine Gruppe von Leuten, die das nächste Jahr feiern.
Explicit Length Constraints

• All models satisfy given length
• Encoding outperforms learned embedding
  - Unseen lengths in training

<table>
<thead>
<tr>
<th>Adapted models</th>
<th>WER ↓</th>
<th>R-1 ↑</th>
<th>R-2 ↑</th>
<th>R-L ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (stop dec.)</td>
<td>39.9</td>
<td>74.6</td>
<td>57.0</td>
<td>72.6</td>
</tr>
<tr>
<td>Length embedding</td>
<td>39.3</td>
<td>74.3</td>
<td>55.2</td>
<td>72.5</td>
</tr>
<tr>
<td>Length encoding</td>
<td><strong>38.6</strong></td>
<td><strong>75.1</strong></td>
<td><strong>56.4</strong></td>
<td><strong>73.2</strong></td>
</tr>
</tbody>
</table>
Low-latency Sequence-to-Sequence Models

• Produce translation shortly after words are spoken
  - Before sentence ends

• Very short context

• Two techniques:
  - Iterative updates
  - Local agreement
Iterative Updates

- Directly output first hypothesis
- If more context is available:
  - Update with better hypothesis

Example:
- Ich melde mich
- I register
- Ich melde mich von der Klausur ab
- I withdraw form the exam

- Not only for MT, but for all components [Niehues et al, 2016]
Adaptation to NMT

- **Challenge:**
  - NMT always tries to generate complete sentence
  - Example:
    - I encourage all of
    - Yo animo a todo el mundo.
Adaptation to NMT

- **Idea:**
  - Train NMT on partial sentences
  - No parallel data available -> Generate artificial data

- **Source data:**
  - Every prefix of the sentence

- **Target data:**
  - **Constraints:**
    - As long as possible for low latency
    - Substring of previous prefix for few rewrites
  - **Length-based**
    - Same ratio of source and target sentence
  - **Alignment-based:**
    - Giza++ alignment
    - Longest prefix that no target word aligned outside source prefix
Adaptation to NMT

- Training
  - Continue training
    - Performance drop on full sentences
  - Multi-task training
    - Mix partial and full sentences
    - Ratio 1:1
Results

![BLEU Graph](image1)

![Word update Graph](image2)

Jan Niehues - Teaching Machine Translation additional Constraints
Constrained Output

- Simulation framework
  - Evaluation different strategies

Ma et al., 2019
Stream decoding strategies

- **Wait-k**
  - Wait for k seconds
  - Then output with fixed rate

<table>
<thead>
<tr>
<th>Chunks</th>
<th>Displayed</th>
<th>Output</th>
<th>Prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ø</td>
<td>All model trains</td>
<td>Ø</td>
</tr>
<tr>
<td>1,2</td>
<td>Ø</td>
<td>All model art</td>
<td>All</td>
</tr>
<tr>
<td>1,2,3</td>
<td>All model</td>
<td>All models are wrong</td>
<td>All model are</td>
</tr>
<tr>
<td>1,2,3,4</td>
<td>All model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Stream decoding strategies

- **Hold-n**
  - Do not output last n tokens

<table>
<thead>
<tr>
<th>Chunks</th>
<th>Displayed</th>
<th>Output</th>
<th>Prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ø</td>
<td>All model trains</td>
<td>All model</td>
</tr>
<tr>
<td>1,2</td>
<td>All model</td>
<td>All model art</td>
<td>All model</td>
</tr>
<tr>
<td>1,2,3</td>
<td>All model</td>
<td>All model are wrong</td>
<td>All model are</td>
</tr>
<tr>
<td>1,2,3,4</td>
<td>All model</td>
<td>All model are</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
### Stream decoding strategies

- **Local agreement**
  - Output if previous and current output agree on prefix

<table>
<thead>
<tr>
<th>Chunks</th>
<th>Displayed</th>
<th>Output</th>
<th>Prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ø</td>
<td>All model trains</td>
<td>Ø</td>
</tr>
<tr>
<td>1,2</td>
<td>Ø</td>
<td>All models art</td>
<td>All</td>
</tr>
<tr>
<td>1,2,3</td>
<td>All</td>
<td>All models are wrong</td>
<td>All models</td>
</tr>
<tr>
<td>1,2,3,4</td>
<td>All models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Latency vs. Accuracy

- Speech recognition results
Adaptation

- Adaptation to partial sentences:
  - Train on full and partial sentences

<table>
<thead>
<tr>
<th></th>
<th>Unidirectional</th>
<th>Bi-directional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offline</td>
<td>14.4</td>
<td>14.9</td>
</tr>
<tr>
<td>Local agreement</td>
<td>16.8</td>
<td>15.8</td>
</tr>
<tr>
<td>+Adapt</td>
<td>15.5</td>
<td>15.8</td>
</tr>
</tbody>
</table>
## Speech Translation

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>Latency diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offline</td>
<td>44.5</td>
<td>4.36</td>
</tr>
<tr>
<td>Hold-2</td>
<td>37.3</td>
<td>0.48</td>
</tr>
<tr>
<td>Hold-4</td>
<td>42.2</td>
<td>0.95</td>
</tr>
<tr>
<td>Local Agreement</td>
<td>42.1</td>
<td>0.71</td>
</tr>
</tbody>
</table>
Conclusion

• Integration of additional constraints in NMT
  - Length-constraints
  - Time-constraints

• Architectural changes

• Pseudo-supervised training

• Length-constraints
  - Compared to human subtitles

• Time-constraints
  - Local agreement
Reference


Thanks