Document-Level Machine Translation with Hierarchical Attention

Master’s Defense by Yu-Tang Shen
Advisor: Dr. Chris Pollett
Committee: Dr. Thomas Austin
Committee: Dr. William Andreopoulos
Agenda

- Project Goals
- Background
- Implementation
- Results
- Demo
- Future Work
Project Goals

Problem statement

Document translation requires too much computing power.
Provide **correlations** during the translation process.
## Background

### Terminologies

<table>
<thead>
<tr>
<th>Machine translation</th>
<th>MT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source language</td>
<td>SL</td>
</tr>
<tr>
<td>Target language</td>
<td>TL</td>
</tr>
<tr>
<td>Hierarchical attention</td>
<td>HAN</td>
</tr>
</tbody>
</table>
Background

Machine translation history

- Rule-based MT
- Statistical MT
- Neural MT

before the Transformer
the Transformer
Background

Rule-based machine translation

○ Rules

<table>
<thead>
<tr>
<th>Mary is</th>
<th>瑪莉是</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>那位</td>
</tr>
<tr>
<td>lady</td>
<td>女士</td>
</tr>
<tr>
<td>in a white dress</td>
<td>穿白裙子的</td>
</tr>
</tbody>
</table>

○ Results

瑪莉是那位穿白裙子的女士
Background
Rule-based machine translation

Pros
○ **Simple** algorithm to implement
○ Deterministic results
  ● Easy to debug

Cons
○ **Labor-intensive** for listing out the rules
○ Deterministic results
  ● Can’t resolve lexical ambiguity
    ■ I saw bats
Background

Statistical machine translation

- Translates based on **language statistics**
- **Flexible** translations

Saw

- Past tense of "see"
- A hand tool for cutting wood
Background

Statistical machine translation

○ Language statistics
  ○ Frequency of SL $s$ being translated TL $t$
    
    | saw     | 看到 (see) | 90% |
    |---------|-----------|-----|
    |         | 錾子 (tool) | 10% |

  ○ Frequency of TL $t_1$ following TL $t_0$
    
    | $P(saw = 看到 (see) | 我 (I))$ | 90% |
    | $P(saw = 錾子 (tool) | 我 (I))$ | 10% |
## Background

### Statistical machine translation

- **Statistics**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>l</td>
<td>我</td>
<td>100%</td>
</tr>
<tr>
<td>saw</td>
<td>看到 (see)</td>
<td>90%</td>
</tr>
<tr>
<td></td>
<td>鋸子 (tool)</td>
<td>10%</td>
</tr>
<tr>
<td>bats</td>
<td>蝙蝠 (animal)</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>球棒 (stick)</td>
<td>50%</td>
</tr>
<tr>
<td>P(saw = 看到 (see)</td>
<td>我)</td>
<td>90%</td>
</tr>
<tr>
<td>P(saw = 鋸子 (tool)</td>
<td>我)</td>
<td>10%</td>
</tr>
<tr>
<td>P(bats = 蝙蝠 (animal)</td>
<td>看到 (see)</td>
<td>55%</td>
</tr>
<tr>
<td>P(bats = 球棒 (stick)</td>
<td>看到 (see)</td>
<td>45%</td>
</tr>
</tbody>
</table>

P(saw = 看到 (see) | band) | 10%

P(saw = 鋸子 (tool) | band) | 90%

P(saw = 鋸子 (tool) | a) | 99%

...
### Background

#### Statistical machine translation

- **Statistics**

<table>
<thead>
<tr>
<th>l</th>
<th>我</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>saw</td>
<td>看到 (see)</td>
<td>90%</td>
</tr>
<tr>
<td></td>
<td>锯子 (tool)</td>
<td>10%</td>
</tr>
<tr>
<td>bats</td>
<td>蝙蝠 (animal)</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>球棒 (stick)</td>
<td>50%</td>
</tr>
</tbody>
</table>

- $P(\text{saw} = \text{看到 (see)} | \text{我}) = 90\%$
- $P(\text{saw} = \text{锯子 (tool)} | \text{我}) = 10\%$
- $P(\text{bats} = \text{蝙蝠 (animal)} | \text{看到 (see)}) = 55\%$
- $P(\text{bats} = \text{球棒 (stick)} | \text{看到 (see)}) = 45\%$

- $\arg\max_x P(l = x)$

- 我
Background

Statistical machine translation

- Statistics

<table>
<thead>
<tr>
<th>l</th>
<th>我</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>saw</td>
<td>看到 (see)</td>
<td>90%</td>
</tr>
<tr>
<td>溝子 (tool)</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>bats</td>
<td>蝙蝠 (animal)</td>
<td>50%</td>
</tr>
<tr>
<td>球棒 (stick)</td>
<td>50%</td>
<td></td>
</tr>
</tbody>
</table>

P(saw = 看到 (see) | 我) | 90%|
P(saw = 溝子 (tool) | 我) | 10%|
P(bats = 蝙蝠 (animal) | 看到 (see) | 55%|
P(bats = 球棒 (stick) | 看到 (see) | 45%|

- argmax_x P(l = x)
  - 我

- argmax_x P(saw = x | 我)
### Background

**Statistical machine translation**

#### Statistics

<table>
<thead>
<tr>
<th>I</th>
<th>我</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>saw</td>
<td>看到 (see)</td>
<td>90%</td>
</tr>
<tr>
<td></td>
<td>鋸子 (tool)</td>
<td>10%</td>
</tr>
<tr>
<td>bats</td>
<td>蝙蝠 (animal)</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>球棒 (stick)</td>
<td>50%</td>
</tr>
</tbody>
</table>

- $P(\text{saw} = \text{看到 (see)} \mid \text{我}) = 90\%$
- $P(\text{saw} = \text{鋸子 (tool)} \mid \text{我}) = 10\%$
- $P(\text{bats} = \text{蝙蝠 (animal)} \mid \text{看到 (see)}) = 55\%$
- $P(\text{bats} = \text{球棒 (stick)} \mid \text{看到 (see)}) = 45\%$

**argmax_x P(I = x)**

- 我

**argmax_x P(saw = x \mid 我)**

- 看到 (see)
Background

Statistical machine translation

- **Statistics**

<table>
<thead>
<tr>
<th></th>
<th>我</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>saw</td>
<td>看到 (see)</td>
<td>90%</td>
</tr>
<tr>
<td></td>
<td>鋸子 (tool)</td>
<td>10%</td>
</tr>
<tr>
<td>bats</td>
<td>蝙蝠 (animal)</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>球棒 (stick)</td>
<td>50%</td>
</tr>
</tbody>
</table>

- argmax_x P (l = x)
  - 我

- argmax_x P (saw = x | 我)
  - 看到 (see)

- argmax_x P (bats = x | 看到 (see))
## Background

### Statistical machine translation

#### Statistics

<table>
<thead>
<tr>
<th></th>
<th>我</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>saw</td>
<td>看到 (see)</td>
<td>90%</td>
</tr>
<tr>
<td>bats</td>
<td>蝙蝠 (animal)</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>鋸子 (tool)</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>球棒 (stick)</td>
<td>50%</td>
</tr>
</tbody>
</table>

- $P(\text{saw} = \text{看到 (see)} | \text{我}) = 90\%$
- $P(\text{saw} = \text{鋸子 (tool)} | \text{我}) = 10\%$
- $P(\text{bats} = \text{蝙蝠 (animal)} | \text{看到 (see)}) = 50\%$
- $P(\text{bats} = \text{球棒 (stick)} | \text{看到 (see)}) = 50\%$

- $\text{argmax}_x P(\text{I} = x)$
  - 我

- $\text{argmax}_x P(\text{saw} = x | \text{我})$
  - 看到 (see)

- $\text{argmax}_x P(\text{bats} = x | \text{看到 (see)})$
  - 蝙蝠 (animal)

- Final result:
  - 我看到 (see) 蝙蝠 (animal)
Background

Statistical machine translation

Pros

○ Flexible translations
  ● Better readability\(^1\)

Cons

○ Doesn’t analyze / interpret the context thoroughly
  ● The band saw some audience leaving the concert early.

\[^1\text{As seen in CS 297 experiments at http://www.cs.sjsu.edu/faculty/pollett/masters/Semesters/Fall22/thomas/index.php?297Deliverable2.php}\]
Background

Neural machine translation

- Neural networks are good at resolving complex correlations
  - Translations are complicated matching
Background
Neural machine translation
Background

Neural machine translation

*Encoder-decoder*
Background

Neural machine translation

Encoder-decoder
Background

Neural machine translation

Encoder-decoder
Background

Neural machine translation

*Interlingua*

- **Direct translation**
  - $SL \rightarrow TL$

- **Indirect translation / transfer translation**
  - $SL \rightarrow \text{Interlingua} \rightarrow TL$
Background

Neural machine translation

*Encoder-decoder*

- **Indirect translation / transfer translation**
  - SL => Interlingua => TL
Background

Neural machine translation

Encoder-decoder

○ Encoder / decoder options
  ● RNN
  ● LSTM
Background

Neural machine translation

The Transformer [1]

○ Each token attends with others
  • $\alpha(\text{“surgeon”}, \text{“surgery”}) > \alpha(\text{“engineer”}, \text{“surgery”})$

○ What is important to translate this token

Background

Neural machine translation

*The Transformer*

- **Pros**
  - Avoid gradient vanishing / exploding
  - More parallel $\Rightarrow$ efficient
Background

Neural machine translation

*Big Bird [2] Attention mechanism*

- Attend to the tokens that are **more likely to be relevant**

Background

Neural machine translation

*Big Bird Attention mechanism*

○ Attend to the tokens that are *more likely to be relevant*
Background

Neural machine translation

*Big Bird Attention mechanism*

- Attend to the tokens that are more likely to be relevant
Background

Neural machine translation

Flaw

- **Encoder input length limit**
Implementation

Model design
Implementation

Overview

- Preprocessing
  - K dimensional tree (k-d tree)
  - The digit issue

- New layers
  - Autoencoder
  - Hierarchical attention (HAN)
    - Big Bird attention

- New model metric
Implementation

Preprocessing

- Store data as vectors instead of words
  - spaCy tokenization
  - Fasttext word vectors
  - Store into TensorFlow Dataset
Implementation

Preprocessing

*K-d tree*

- Decipher word vector outputs
- K-d tree is a fast algorithm ($O(\log n)$) for spatial search
Implementation

Preprocessing

$K$-d tree

Diagram (a) shows a set of points in a 2D space, represented by circles labeled A, B, C, D, E, F, and G. These points are plotted on an x-y axis. Diagram (b) illustrates a K-d tree structure built from these points, with nodes A, B, E, D, F, and G. Diagram (c) presents a dictionary with key-value pairs to associate with the points in diagram (a):

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1, 2]</td>
<td>D</td>
</tr>
<tr>
<td>[2, 6]</td>
<td>C</td>
</tr>
<tr>
<td>[3, 4]</td>
<td>B</td>
</tr>
<tr>
<td>[4, 3]</td>
<td>A</td>
</tr>
<tr>
<td>[5, 1]</td>
<td>G</td>
</tr>
<tr>
<td>[6, 5]</td>
<td>E</td>
</tr>
<tr>
<td>[7, 7]</td>
<td>F</td>
</tr>
</tbody>
</table>
Implementation

Preprocessing

*K-d tree*
Implementation

Preprocessing

The digit issue

- Fail to translate digits
  - Guess
    - Occurrence of 2008 was too little
  - Solution
    - Split 2008 into 2, 0, 0, and 8

```python
1  zh_devectorize(result)

[START]柏林—2008年爆发的全球金融和经济危机是自大萧条以来最严重
的管理和治理缺陷。事实上，2008年危机极有可能被视
为预防政策引发未来几十年来新的经济
和金融体系。尽管这一目标并不全无价值，但就像历史学家们
discussion.

1  zh_devectorize(zh[0])

柏林—2011年爆发的全球金融和经济危机是自大萧条以来最严重

| 监管机制，仅仅是事实上地毫无疑问金融金融体系。尽管这一技术进步所带来的挑战，就必须对国内和国际两
| 此时新成立的金融金融体系将会进一步与日俱增，这些举措最终会增加
| 技术进步，并进一步加大金融业务的监督和金融体系。尽管这一目标并不全无价值，但就像历史学家们
discussion.
```
Implementation
Model design
Implementation

New layers

Autoencoder

- Autoencoder summarizes the input on a sentence level* to get context information

Output shape: \([\text{batch\_size}, \text{num\_sent}^*, \text{sent\_embedding\_size}]\)

Input shape: \([\text{batch\_size}, \text{num\_tokens}, \text{token\_embedding\_size}]\)
Implementation

New layers

Autoencoder

- Summarizing actual sentences is **slow**
- Set a **fixed size window and stride** to split input into sentences
  - Window: 16 (average sentence length in dataset = 14)
  - Stride: 8
Implementation

New layers

\textit{HAN}

- Each token attends with
  - Other tokens, and
  - Context information nodes

- Information
  - At that position, and
  - From distant nodes
Implementation

New model metric

- Penalize by **distance**
  - *Mean squared error* as loss function

- Evaluate with **hit or miss**
  - Hit if a dimension of output vector is within a defined threshold
  - Miss if otherwise
  - *Accuracy = #hits / total dimensions*

<table>
<thead>
<tr>
<th>Truth</th>
<th>Prediction</th>
<th>Absolute difference</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>[10, 10], [20, 20]</td>
<td>[11, 23], [19, 24]</td>
<td>[1, 13], [1, 4]</td>
<td>5</td>
</tr>
</tbody>
</table>

Accuracy = 3 / 4
Results

Overview

- English to Chinese translation
  - Full attention
  - Big Bird attention

- Chinese to English translation
  - Big Bird attention
**Results**

**English to Chinese translation**

*Full attention*

<table>
<thead>
<tr>
<th>Input</th>
<th>Prediction</th>
<th>Ground truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERLIN – The <strong>global</strong> financial and <strong>economical</strong> crisis that began in 2008 was the greatest <strong>economic</strong> stress-test since the Great Depression, and the greatest <strong>challenge</strong> to <strong>social</strong> and <strong>political</strong> systems since World War II.</td>
<td>柏林——2008年的全球金融经济经济危机自大萧条以来最严峻的一次经济经济压力测试测试，是自自以来社会社会政治政治所面临的最严重挑战。</td>
<td>柏林——2008年爆发的全球金融和经济危机是自大萧条以来最严峻的一次经济压力测试，也是自二战以来社会和政治制度所面临的最严重挑战。</td>
</tr>
</tbody>
</table>
Results

English to Chinese translation

*Big Bird attention*
## Results

### English to Chinese translation

**Big Bird attention**

<table>
<thead>
<tr>
<th>Input</th>
<th>Prediction</th>
<th>Ground truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERLIN – The global financial and economical crisis that began in 2008 was the greatest economic stress-test since the Great Depression, and the greatest challenge to social and political systems since World War II.</td>
<td>柏林——2008年爆发的全球金融和经济危机是自大萧条以来最严峻的一次经济压力测试，也是自二战以来社会和政治制度所面临的最严重挑战。</td>
<td>柏林——2008年爆发的全球金融和经济危机是自大萧条以来最严峻的一次经济压力测试，也是自二战以来社会和政治制度所面临的最严重挑战</td>
</tr>
</tbody>
</table>
Results

Chinese to English translation

*Big Bird attention*

- **Model complexity reduced** due to larger input space
  - Chinese documents are longer
  - Number of attention layers: 4 => 2
  - Number of attention heads: 8 => 4
### Results

#### Chinese to English translation

*Big Bird attention*

<table>
<thead>
<tr>
<th>Input</th>
<th>Prediction</th>
<th>Ground truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>柏林——2008年爆发的全球经济和经济危机是自大萧条以来最严峻的一次经济压力测试，也是自二战以来社会和政治制度所面临的最严重挑战。</td>
<td><strong>Amsterdam</strong> – the global financial and economic crisis that began in 2008 was the greatest economic stress-test since the Great Depression, and the greatest challenge to social and political systems since World War II.</td>
<td><strong>BERLIN</strong> – The global financial and economical crisis that began in 2008 was the greatest economic stress-test since the Great Depression, and the greatest challenge to social and political systems since World War II.</td>
</tr>
</tbody>
</table>
## Results

### Comparison

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Configuration</th>
<th>Results</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Attention layers</td>
<td>Attention heads</td>
<td>Validation Accuracy</td>
<td>BLEU score</td>
<td>Training cost (ms/step)</td>
</tr>
<tr>
<td>HAN</td>
<td>4</td>
<td>8</td>
<td>0.8</td>
<td>0.44</td>
<td>172</td>
</tr>
<tr>
<td>HAN-SD</td>
<td>4</td>
<td>8</td>
<td>0.92</td>
<td>0.9</td>
<td>169</td>
</tr>
<tr>
<td>BB-HAN-EN_ZH</td>
<td>4</td>
<td>8</td>
<td>0.96</td>
<td>0.86</td>
<td>171</td>
</tr>
<tr>
<td>BB-HAN-ZH_EN</td>
<td>2</td>
<td>4</td>
<td>0.95</td>
<td>0.79</td>
<td>81</td>
</tr>
</tbody>
</table>
Results

Observation

Desirable feature

○ Comprehendible texts

Bugs

○ Stuttering effect
○ Smaller models may cause incorrect translations
Demo
4 min
[ ]
1. VALIDATION_SIZE = 0.3
2. MAX_TOKENS = 4096
3. CHUNK_SIZE = 16 # English sentence average sentence length: 15~20 / Chinese sentence: 8~14
4. LATENT_SIZE = 300
5. BBRANDOM_RATIO = 0.3
6. BATCH_SIZE = 4
7. THRESHOLD = 0.05

Show code

[ ]
1. from google.colab import drive
2. drive.mount('/content/drive')
3. %cd drive/MyDrive/HAN

Mounted at /content/drive

56s completed at 3:45 AM
Future work

- Optimize word vectors during training
- Post editing
  - Removing stuttering effect in English
- Optimize Big Bird attention mechanism to boost efficiency
Thank you!
Attention Q, K, V?

- Input sequence [1, 2, 3]
  - 3x3 matrix

- Q => **input * 2 = [2, 4, 6]**
  - Q => input * \( W_q \)
Multihead attention

Add & Norm

FF

Add & Norm

q k v q k v q k v

q k v q k v q k v

q k v q k v q k v
Multihead attention
## Implementation

### New model metric

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>$[y_0, y_1, y_2, \ldots, y_{\text{wordsInOutput}}]$</td>
<td>$[v_0, v_1, v_2, \ldots, v_{\text{wordsInTruth}}]$</td>
</tr>
<tr>
<td>$y_0 = [y_0^0, y_0^1, y_0^2, \ldots, y_0^{299}]$</td>
<td>$v_0 = [v_0^0, v_0^1, \ldots, v_0^{299}]$</td>
</tr>
</tbody>
</table>

### Distance between $y_0$ and $v_0$

$$\sqrt{(y - v)^2} = \sqrt{(y_0^0 - v_0^0)^2 + \cdots + (y_0^{299} - v_0^{299})^2}$$

$$= \sqrt{\dim(y) \sum_{i=0}^{\dim(y)} (y_0^i - v_0^i)^2}$$

### MSE between $y_0$ and $v_0$

$$\frac{1}{\dim(y)} \sum_{i=0}^{\dim(y)} (y_0^i - v_0^i)^2$$

$$\sqrt{\sum_{i=0}^{\dim(y)} (y_0^i - v_0^i)^2} \propto \frac{1}{\dim(y)} \sum_{i=0}^{\dim(y)} (y_0^i - v_0^i)^2$$