Machine Translation at Booking.com Journey and Lessons Learned

May 30, 2017, Prague

Pavel Levin
Nishikant Dhanuka
Maxim Khalilov
Who am I?

About me

- Master in Computer Science (NLP) from IIT Mumbai
- 8 years of work experience in analytics and consulting
- Data Scientist at Booking.com since last 2 years
  - Partner Services Department (Scaled Content)
- linkedin.com/in/nishikantdhanuka/

About Booking.com

World’s #1 website for booking hotels and other accommodations

- Founded in 1996 in Amsterdam; part of The Priceline Group since 2005
- 1,200,000+ properties in more than 220+ countries; 25 Million rooms
- Over 1,400,000 room nights are reserved on Booking.com every 24 hours
- Employing more than 15,000 people in 199 offices worldwide
- Website available in 43 languages
Agenda.

• **Motivation**
  ➡ MT critical for Booking.com’s localization process

• **MT Journey and Lessons Learned**
  ➡ MT Model & Experiments
  ➡ Evaluation Results
    ◦ Automatic & Human
    ◦ Sentence Length Analysis
    ◦ A/B Tests
  ➡ Interesting Examples

• **Conclusion and Future Work**
• **Q & A**
Motivation
**Mission:** Empower people to book any hotel in the world, while browsing high quality content in their own language.

2/3 of daily bookings on Booking.com is made in a language other than English.

... thus it is important to have **locally relevant content at scale**

### How Locally Relevant?
- Present **Hotel descriptions** in the language of the user
- Allow partners and guests to **consume and produce content in their own language**
  - Customer Reviews
  - Customer Service Support

### Why At Scale?
- **One Million+ properties** and growing very fast
- **Frequent change requests** to update the content
- **43 languages** and more
- New **customer reviews / tickets** every second
Currently Hotel descriptions translated by human in 43 languages based on visitor demand.

*50% Translation Coverage

*90% Demand Coverage

* Approximate numbers based on average of some languages
Example of Lost Business Opportunity because of highly manual and slow process.

Put in human translation pipeline if this happens often

New Hotel in China
Initial content only in English & Chinese

Profile visited on B.com by a German customer

Sees the description in English

Drops Off

Still makes the booking

Success

Lost Business

How do we balance quality, speed and cost effectiveness?

Chicken-Egg problem

Machine Translation

Booking.com
MT Journey and Lessons Learned
Our Journey to discover the awesomeness of NMT

General Purpose
Trained on general purpose data

Booking.com
Trained on in-domain data

Phase 1
SMT

Phase 2
NMT

Phase 3
NMT

In-domain SMT
> General Purpose SMT

General Purpose NMT
> In-domain SMT

In-domain NMT
> General Purpose NMT
Lots of in-domain data to train the MT system

<table>
<thead>
<tr>
<th>Language</th>
<th>Parallel Sentences</th>
<th># of Words</th>
<th>Vocab Size</th>
<th>Avg. Len</th>
</tr>
</thead>
<tbody>
<tr>
<td>English -&gt; German</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>10.5 M</td>
<td>171 M</td>
<td>845 K</td>
<td>16.3</td>
</tr>
<tr>
<td>English</td>
<td>174 M</td>
<td>583 K</td>
<td></td>
<td>16.5</td>
</tr>
<tr>
<td>English -&gt; French</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>11.3 M</td>
<td>193 M</td>
<td>588 K</td>
<td>17.7</td>
</tr>
<tr>
<td>English</td>
<td>188 M</td>
<td>581 K</td>
<td></td>
<td>16.7</td>
</tr>
</tbody>
</table>
## Our NMT Model Configuration Details

<table>
<thead>
<tr>
<th>Data Preparation</th>
<th>Model</th>
<th>Training</th>
<th>Translate</th>
<th>Evaluate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Split Data</strong></td>
<td><strong>Model Type</strong></td>
<td><strong>Optimization Method</strong></td>
<td><strong>Beam Size</strong></td>
<td><strong>Auto</strong></td>
</tr>
<tr>
<td>Train, Val, Test</td>
<td>seq2seq</td>
<td>Stochastic Gradient Descent</td>
<td>30</td>
<td>BLEU WER</td>
</tr>
<tr>
<td><strong>Input Text Unit</strong></td>
<td><strong>Input Embedding Dimension</strong></td>
<td><strong>Initial Learning Rate</strong></td>
<td><strong>Unknown Words Handling</strong></td>
<td>Human A/F</td>
</tr>
<tr>
<td>Word Level</td>
<td>1,000</td>
<td>1</td>
<td>Source with Highest Attention</td>
<td></td>
</tr>
<tr>
<td><strong>Tokenization</strong></td>
<td><strong>RNN Type</strong></td>
<td><strong>Decay Rate</strong></td>
<td></td>
<td><strong>Other</strong></td>
</tr>
<tr>
<td>Aggressive</td>
<td>LSTM</td>
<td>0.5</td>
<td></td>
<td>Length A/B Test</td>
</tr>
<tr>
<td><strong>Max Sentence Length</strong></td>
<td><strong># of hidden layers</strong></td>
<td><strong>Decay Strategy</strong></td>
<td></td>
<td><strong>MT pipeline based on Harvard implementation of OpenNMT</strong></td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>Decrease in Validation Perplexity &lt;=0</td>
<td></td>
<td><strong>1 Epoch takes approx. 2 days on a single NVIDIA Tesla K80 GPU</strong></td>
</tr>
<tr>
<td><strong>Vocabulary Size</strong></td>
<td><strong>Hidden Layer Dimension</strong></td>
<td><strong>Number of Epochs</strong></td>
<td></td>
<td><strong>Approx. 220 Million Parameters</strong></td>
</tr>
<tr>
<td>50,000</td>
<td>1,000</td>
<td>5 - 13</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Attention Mechanism</strong></td>
<td><strong>Stopping Criteria</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Global Attention</td>
<td>BLEU + sensitive sentences + constraints</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Initial Parameters</strong></td>
<td><strong>Dropout</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Batch Size</strong></td>
<td>0.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>250</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:**
- **Approx. 220 Million Parameters**
- **1 Epoch takes approx. 2 days on a single NVIDIA Tesla K80 GPU**
1. Data Preparation: Tokenization and Vocabulary

<table>
<thead>
<tr>
<th>Data Preparation</th>
<th>EN: The rooms at the Prague Mandarin Oriental feature underfloor heating, and guests can choose from various bed linen and pillows.</th>
<th>DE: Die Zimmer im Prague Mandarin Oriental bieten eine Fußbodenheizung und eine Auswahl an Bettwäsche und Kissen.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Split Data</strong></td>
<td>Train, Val, Test</td>
<td></td>
</tr>
<tr>
<td><strong>Input Text Unit</strong></td>
<td>Word Level</td>
<td></td>
</tr>
<tr>
<td><strong>Tokenization</strong></td>
<td>Aggressive</td>
<td></td>
</tr>
<tr>
<td><strong>Max Sentence Length</strong></td>
<td>50</td>
<td></td>
</tr>
<tr>
<td><strong>Vocabulary Size</strong></td>
<td>50,000</td>
<td></td>
</tr>
</tbody>
</table>

*Aggressive* only keeps sequences of letters / numbers i.e. doesn’t allow mix of alphanumeric as in: "E65", "soft-landing"

Tokenized text represented as bag of words vector based on vocabulary ids.
2. Model Architecture: **Approx. 220 Million Parameters**

<table>
<thead>
<tr>
<th>Model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model Type</strong></td>
<td>seq2seq</td>
</tr>
<tr>
<td><strong>Input Embedding Dimension</strong></td>
<td>1,000</td>
</tr>
<tr>
<td><strong>RNN Type</strong></td>
<td>LSTM</td>
</tr>
<tr>
<td><strong># of hidden layers</strong></td>
<td>4</td>
</tr>
<tr>
<td><strong>Hidden Layer Dimension</strong></td>
<td>1,000</td>
</tr>
<tr>
<td><strong>Attention Mechanism</strong></td>
<td>Global Attention</td>
</tr>
</tbody>
</table>

Inclives **Wifi**.
3. Training: 1 Epoch takes approx. 2 days on a single NVIDIA Tesla K80 GPU

<table>
<thead>
<tr>
<th>Training</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimization Method</td>
<td>Stochastic Gradient Descent</td>
</tr>
<tr>
<td>Initial Learning Rate</td>
<td>1</td>
</tr>
<tr>
<td>Decay Rate</td>
<td>0.5</td>
</tr>
<tr>
<td>Decay Strategy</td>
<td>Decrease in Validation Perplexity &lt;=0</td>
</tr>
<tr>
<td>Number of Epochs</td>
<td>5 - 13</td>
</tr>
<tr>
<td>Stopping Criteria</td>
<td>BLEU + sensitive sentences +constraints</td>
</tr>
<tr>
<td>Dropout</td>
<td>0.3</td>
</tr>
<tr>
<td>Batch Size</td>
<td>250</td>
</tr>
</tbody>
</table>

**Stopping Criteria:** Sensitive Sentence Example

*The neighborhood is very nice and safe*

*There is a safe installed in this very nice neighborhood*
### 4. Translate: Unknown Word Handling

<table>
<thead>
<tr>
<th>Translate</th>
<th>Good Example</th>
<th>Bad Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Beam Size</strong></td>
<td>Offering a restaurant, Hodor Eco-lodge is located in Winterfell.</td>
<td>Free access to The Game entertainment Centre</td>
</tr>
<tr>
<td><strong>Unknown Words Handling</strong></td>
<td><strong>Source</strong></td>
<td><strong>Human Translation</strong></td>
</tr>
<tr>
<td></td>
<td>Source with Highest Attention</td>
<td>Das Hodor Eco-Lodge begrüßt Sie in Winterfell mit einem Restaurant.</td>
</tr>
<tr>
<td></td>
<td><strong>Human Translation</strong></td>
<td>Kostenlosfreier Zugang zum Unterhaltungszentrum The Game</td>
</tr>
<tr>
<td></td>
<td>Raw Output</td>
<td>Output with &lt;unk&gt; replaced</td>
</tr>
<tr>
<td></td>
<td>Das &lt;unk&gt;&lt;unk&gt; in &lt;unk&gt; bietet ein Restaurant.</td>
<td>Das Hodor Eco-lodge in Winterfell bietet ein Restaurant.</td>
</tr>
<tr>
<td></td>
<td><strong>Output with &lt;unk&gt; replaced</strong></td>
<td>Kostenlosfreier Zugang zum Centre</td>
</tr>
</tbody>
</table>
5. Evaluate: **Auto, Human, Length Analysis & A/B Tests**

<table>
<thead>
<tr>
<th>Evaluate</th>
<th>BLEU</th>
<th>A/F Framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto</td>
<td>BLEU WER</td>
<td>3 evaluators per language</td>
</tr>
<tr>
<td>Human</td>
<td>A/F</td>
<td>Provided with original text and MT hypotheses, including human reference</td>
</tr>
<tr>
<td>Other</td>
<td>Length A/B Test</td>
<td>Not aware which system produced which hypothesis</td>
</tr>
</tbody>
</table>

**BLEU**
- # of words shared between MT output and human reference
  - Benefits sequential words
  - Penalizes short translations

**WER**
- Variation of the word-level Levenshtein distance
  - Measures the distance by counting insertions, deletions & substitutions

**A/F Framework**
- Asked to assess the quality of 150 random sentences from test corpus
- 4 level scale to both Adequacy & Fluency

**Example:**

**Minor Mistake:**
- EN: “there is a parking area available”
- DE: “es ist eine Garage verfügbar”

**Major Mistake:**
- EN: “there is a parking area available”
- DE: “es ist eine Aufbewahrungsstelle verfügbar”
Evaluation Results 1/5:
BLEU Score for German & French

Our In-domain NMT system significantly outperforms all other MT engines.
Both Neural systems consistently outperform their Statistical counter-parts.
In-domain SMT beats General Purpose NMT.
Compared to German, French improved much more from SMT to NMT.
Evaluation Results 2/5:
Adequacy/Fluency Scores for German

- Adequacy:
  - SMT: 3.62
  - NMT: 3.9
  - GP-SMT: 3.57
  - GP-NMT: 3.65
  - Human: 3.96

- Fluency:
  - SMT: 3.15
  - NMT: 3.78
  - GP-SMT: 3.37
  - GP-NMT: 3.57
  - Human: 3.82

Our In-domain NMT system still outperforms all other MT engines.

Both Neural systems still consistently outperform their statistical counter-parts.

However General Purpose NMT now beats In-domain SMT.

Particularly fluency score of our NMT engine is close to human level.
Evaluation Results 3/5: Adequacy/Fluency Scores for French

Adequacy Scores:
- Human: 3.75
- SMT: 3.4
- NMT: 3.67
- GP-SMT: 3.32
- GP-NMT: 3.78

Fluency Scores:
- Human: 3.75
- SMT: 3.28
- NMT: 3.4
- GP-SMT: 3.31
- GP-NMT: 3.41

General Purpose NMT system outperforms others; conflicts with BLEU

Apparently General Purpose NMT even outperforms human level

Adequacy of both neural engines is almost at human level; fluency very far though

Compared to German, A/F scores are relatively less for French; conflicts with BLEU
Evaluation Results 4/5: BLEU by Sentence Length for German and French

For longer sentences, though performance degraded, NMT still outperformed SMT. Initially performance increases with length, but reaches a peak soon & starts to decline then. For sentences longer than 27 tokens, NMT quality degrades faster than SMT.
Evaluation Results 5/5:
Minus WER by Sentence Length for German and French

For longer sentences, though performance degraded, NMT still outperformed SMT. Initially performance increases with length, but reaches a peak soon & starts to decline then. For sentences longer than 27 tokens, NMT quality degrades faster than SMT.
A/B Tests to validate the hypothesis that MT’ed hotel descriptions have higher conversion than no translation

50% see base with no translation

Offering free WiFi and a garden, VSG Apartment Petrska is situated in Prague, 900 metres from Old Town Square. Prague Astronomical Clock is 1 km away.

The accommodation comes with a seating and dining area. All units feature a kitchen equipped with a dishwasher and microwave. A fridge and coffee machine are also provided. Towels and bed linen are offered.

Wenceslas Square is 1.2 km from VSG Apartment Petrska. The nearest airport is Vaclav Havel Prague Airport, 12 km from the property.

50% see variant with machine translation

Mit kostenfreiem WLAN und einem Garten erwartet Sie das VSG Apartment Petrska in Prag, 900 m vom Altstädter Ring entfernt. Die Astronomische Uhr von Prag erreichen Sie nach 1 km.


Der Wenzelsplatz liegt 1.2 km vom VSG Apartment Petrska entfernt. Der nächste Flughafen ist der 12 km von der Unterkunft entfernte Flughafen Prag.
### Few Interesting Examples from French translations

<table>
<thead>
<tr>
<th>Source</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Good</strong></td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td>The neighbourhood is very safe.</td>
</tr>
<tr>
<td>Word Sense Disambiguation</td>
<td>There is a safe installed in the room.</td>
</tr>
<tr>
<td><strong>Bad</strong></td>
<td></td>
</tr>
<tr>
<td>Bad</td>
<td>The owners are super right wing.</td>
</tr>
<tr>
<td>Out of domain sentence</td>
<td></td>
</tr>
<tr>
<td>Ugly</td>
<td></td>
</tr>
<tr>
<td>Ugly</td>
<td>Sdfdlfsldk offers free breakfast</td>
</tr>
<tr>
<td>OOV words</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion & Future Work
Conclusion.

- NMT consistently outperforms SMT
- In case of German, in-house NMT is also better than online general purpose engines in our application
- Fluency of NMT is close to human translation level
- In our application the relative performance of NMT against SMT does not degrade with increased sentence length
Future Work.

• Improve Unknown Word Handling
  ▶ Explore open vocabulary techniques for UNK handling; sub-word tokenization using byte pair encoding

• Identify Business Sensitive Translation Errors
  ▶ ‘free’ being translated to ‘available’

• Expand to other languages for hotel descriptions
  ▶ Particularly Asian languages like Chinese, Japanese etc.

• Expand to other use cases for MT at Booking.com
  ▶ User generated content like customer reviews, messages etc.
Thank You
Questions?

We’re hiring!
workingatbooking.com

Nishikant Dhanuka from Booking.com
linkedin.com/in/nishikantdhanuka/