70 Years of Machine Translation
Kilimanjaro is a snow-covered mountain 19,710 feet high, and is said to be the highest mountain in Africa. Its western summit is called the Masai “Ngaje Ngai,” the House of God. Close to the western summit there is the dried and frozen carcass of a leopard. No one has explained what the leopard was seeking at that altitude.
Kilimanjaro is a snow-covered mountain 19,710 feet high, and is said to be the highest mountain in Africa. Its western summit is called the Masai “Ngaje Ngai,” the House of God. Close to the western summit there is the dried and frozen carcass of a leopard. No one has explained what the leopard was seeking at that altitude.

Back translation from Japanese (old)
Kilimanjaro is 19,710 feet of the mountain covered with snow, and it is said that the highest mountain in Africa. Top of the west, “Ngaje Ngai” in the Maasai language, has been referred to as the house of God. The top close to the west, there is a dry, frozen carcass of a leopard. Whether the leopard had what the demand at that altitude, there is no that nobody explained.
Kilimanjaro is a snow-covered mountain 19,710 feet high, and is said to be the highest mountain in Africa. Its western summit is called the Masai “Ngaje Ngai,” the House of God. Close to the western summit there is the dried and frozen carcass of a leopard. No one has explained what the leopard was seeking at that altitude.

Back translation from Japanese (old)
Kilimanjaro is 19,710 feet of the mountain covered with snow, and it is said that the highest mountain in Africa. Top of the west, “Ngaje Ngai” in the Maasai language, has been referred to as the house of God. The top close to the west, there is a dry, frozen carcass of a leopard. Whether the leopard had what the demand at that altitude, there is no that nobody explained.

Back translation from Japanese (new)
Kilimanjaro is a mountain of 19,710 feet covered with snow, which is said to be the highest mountain in Africa. The summit of the west is called “Ngaje Ngai” God’s house in Masai language. There is a dried and frozen carcass of a leopard near the summit of the west. No one can explain what the leopard was seeking at that altitude.
Silent launch in Japan...

<table>
<thead>
<tr>
<th>4 hours ago</th>
<th>5 hours ago</th>
<th>6 hours ago</th>
<th>7 hours ago</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google翻訳</td>
<td>Google翻訳</td>
<td>紅の豚</td>
<td>#勇者ヨシヒコ</td>
</tr>
<tr>
<td>#三四郎ann0</td>
<td>#bananamoon</td>
<td>#勇者ヨシヒコ</td>
<td>#bananamoon</td>
</tr>
<tr>
<td>#bananamoon</td>
<td>#勇者ヨシヒコ</td>
<td>#NAOMIの部屋</td>
<td>#勇者ヨシヒコ</td>
</tr>
<tr>
<td>#勇者ヨシヒコ</td>
<td>#NAOMIの部屋</td>
<td>#NAOMIの部屋</td>
<td>#jojo_anime</td>
</tr>
<tr>
<td>#三四郎ann0</td>
<td>#Lostorage</td>
<td>#Lostorage</td>
<td>#Lostorage</td>
</tr>
<tr>
<td>上原甚次</td>
<td>ラブゲ</td>
<td>ラブゲ</td>
<td>ラブゲ</td>
</tr>
<tr>
<td>二人セゾン</td>
<td>二人セゾン</td>
<td>貓の恩返し</td>
<td>#ここち星</td>
</tr>
<tr>
<td>Taka</td>
<td>Taka</td>
<td>砂の塔</td>
<td>#bananamoon</td>
</tr>
<tr>
<td>予想的中</td>
<td>猫の恩返し</td>
<td>猫の恩返し</td>
<td>猫の恩返し</td>
</tr>
</tbody>
</table>

(November 2016)
Quality improvements

- Asian languages improved the most
- Some improvements as big as last 10 years of improvements combined

![Graph showing quality improvements](image)

- 6 = Perfect translation
- 0 = Worst translation
- \( \Delta \) Translation Quality

- Significant change & launchable: \(<0.1\)
- Chinese to English: \(+0.6\)
- Almost all language pairs: \(>0.5\)
- Zh/Ja/Ko/Tr to English: \(0.6-1.5\)
Relative improvement

The diagram compares the translation quality of different models against a perfect translation. The translation models include human, neural (GNMT), and phrase-based (PBMT). The quality is measured on a scale from 0 to 6.
Does quality matter?

+75%

Increase in daily English - Korean translations on Android over the past six months
Old: Phrase-based translation

- Lots of individual pieces
- Optimized somewhat independently

New: Neural machine translation

- End-to-end learning
- Simpler architecture
- Plus results are much better!
Outline

Recent Results

Brief History

Word-based translation

Neural translation

What’s next
Brief history of MT

1954: IBM translates 49 Russian sentences with a 250-word dictionary and 6 grammar rules.

Thomas J Watson “I see in this an instrument that will be helpful in working out the problems of world peace...”
Brief history of MT

1966: Alpac publishes a report concluding that years of research haven't produced useful results. Federal funding for MT research dries up... “AI winter”.

Yehoshua Bar Hillel “The unreasonableness of aiming at fully-automatic high-quality translation is stressed...”
Brief history of MT

1988: IBM Model 1 based on parallel corpora and simple statistical models revives MT.

Peter Brown, Robert Mercer, et al.

The Mathematics of Machine Translation: Parameter Estimation

Peter F. Brown, Stephen A. Della Pietra
Vincent J. Della Pietra, Robert L. Mercer

The availability of large, bilingual corpora has stimulated recent interest in algorithms for manipulating them. A number of authors have discussed algorithms for extracting from such corpora pairs of sentences that are translations of one another. In the course of our work on machine translation, we have developed a series of five statistical models of the translation process. Here, we describe these models and show that it is possible to estimate their parameters automatically from a large set of pairs of sentences. We show, further, that it is possible to align the words within pairs of sentences algorithmically. We have a great deal of data in French and English from the proceedings of the Canadian Parliament. For this reason we have restricted our work to these two languages, but we feel that because our algorithms have minimal linguistic content they would work well on other pairs of languages. We also feel, again because of the minimal linguistic content of our algorithms, that it is reasonable to argue that word-by-word alignments are inherent in any sufficiently large bilingual corpora.
Brief history of MT

1995-2014

- Words -> Phrases
- Scaling up
- A lot of tuning
Outline

Recent Results

Brief History

Word-based translation

Neural translation

What’s next
Translation probabilities

- How to translate a word? Look it up!
  - **Haus** -> house, building, home, household, shell
  - But some more frequent than others...
- We really want to estimate translation probabilities

| e          | $t(e|f)$ |
|------------|---------|
| house      | 0.8     |
| building   | 0.16    |
| home       | 0.02    |
| household  | 0.015   |
| shell      | 0.005   |
Alignments

- Supposing we have parallel text...
- If we had alignments, we could count translations -> estimate probabilities
Alignments

- Supposing we have parallel text...
- If we had alignments, we could count translations -> estimate probabilities
- Alignments come in different flavors
Alignments are latent

- Classic chicken and egg problem!
- If we had alignments, we could estimate translation probabilities...
- ...and if we had translation probabilities, we could generate alignments
- **Solution**: Expectation Maximization Algorithm
EM for Model 1

- Initially, assume all alignments are equally likely (E-step 0)
- Estimate translation probabilities using the alignments (M-step 0)
EM for Model 1

- \( P(\text{the}|\text{la}) > P(\text{house}|\text{la}) \)
- Now, produce new alignments using the updated translation model (E-step 1)
- Again, re-estimate translation model (M-step 1)
EM for Model 1

... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

- After another iteration, *fleur* is aligned to *flower*.
EM for Model 1

- Repeat until convergence
- Note: convergence guaranteed!

\[
p(\text{la} | \text{the}) = 0.453 \\
p(\text{le} | \text{the}) = 0.334 \\
p(\text{maison} | \text{house}) = 0.876 \\
p(\text{bleu} | \text{blue}) = 0.563 \\
\ldots
\]
Decoding

la maison bleu

?
Decoding

- Use the translation model to generate some hypotheses
  - There are a lot of possibilities

```
\begin{align*}
  p(\text{la}|\text{the}) &= 0.453 \\
  p(\text{le}|\text{the}) &= 0.334 \\
  p(\text{maison}|\text{house}) &= 0.876 \\
  p(\text{bleu}|\text{blue}) &= 0.563 \\
  \cdots
\end{align*}
```
Decoding

- Use the translation model to generate some hypotheses
  - There are a lot of possibilities
- Rescore them with a language model
  - This is the noisy channel setup
Decoding

- Use the translation model to generate some hypotheses
  - There are a lot of possibilities
- Rescore them with a language model
  - This is the noisy channel setup
- Beam search
  - Maintain a fixed size stack of partial hypotheses

\[
\begin{align*}
\text{TM:} & \quad p(\text{la | the}) = 0.453 \\
& \quad p(\text{le | the}) = 0.334 \\
& \quad p(\text{maison | house}) = 0.876 \\
& \quad p(\text{bleu | blue}) = 0.563 \\
& \quad \ldots
\end{align*}
\]

\[
\begin{align*}
\text{LM:} & \quad P(\text{house | the}) \\
& \quad P(\text{house | blue}) \\
& \quad P(\text{blue | the}) \\
& \quad \ldots
\end{align*}
\]
IBM Models -> Phrase translation

- Model 1: Lexical translation
- Model 2: Adds absolute reordering model
- Model 3: Adds fertility model
- Model 4: Relative reordering
- Model 5 (and 6): Fix deficiencies
- Phrase translation
Outline

Recent Results
Brief History
Word-based translation
Neural translation
What’s next
2014: Sequence to Sequence

*Sequence to Sequence Learning with Neural Networks* -- Sutskever, Vinyals, Le

- No alignments!
- No language model!
Encoder/Decoder

Encoder/Decoder Recurrent Neural Nets

- Learn to map: X1, X2, EOS -> Y1, Y2, Y3, EOS
- In principle, any lengths should work
- RNN -> LSTM
Deep Architecture
2014: Attention Mechanism

*Neural Machine Translation by Jointly Learning to Align and Translate* -- Bahdanau, Cho, Bengio

- Give decoder access to all encoder states
- Now quality independent of sentence length
Attention Mechanism

- Also, we can retrieve approximate alignments from the attention weights.
Outline

Recent Results
Brief History
Word-based translation
Neural translation
What’s next
Multilingual Models

- Model several language pairs in single model
- Prepend source with additional token to indicate target language
  
  - Translate to Spanish:
    - \(<2es> How are you </s> \rightarrow Cómo estás </s>\)
  
  - Translate to English:
    - \(<2en> Cómo estás </s> \rightarrow How are you </s>\)
- No other changes to model architecture!
The Interlingua

The stratosphere extends from about 10km to about 50km in altitude.

ENGLISH

KOREAN

JAPANESE

성층권은 고도 약 10km부터 약 50km까지 확장됩니다.

成層圏は、高度 10km から 50km の範囲にあります.
What’s next?

- Debugging is hard
- Full document translation
- Use more training data
- More efficient models (avoid RNNs entirely?)
- Translation -> other language tasks