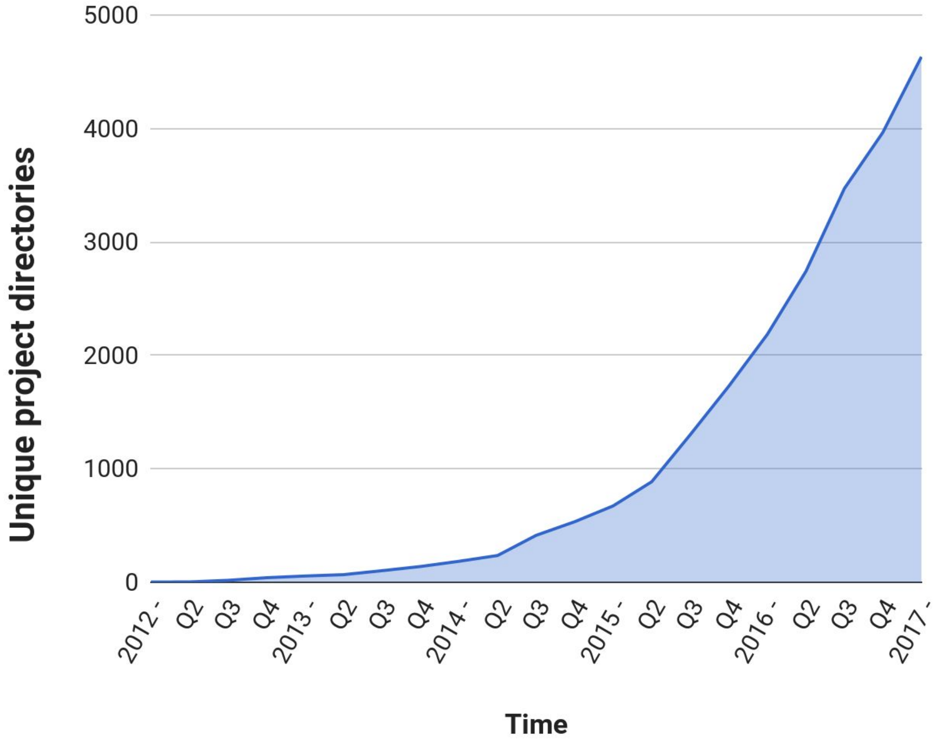
Moving to Neural Machine Translation at Google

Mike Schuster, Google Brain Team

12/18/2017

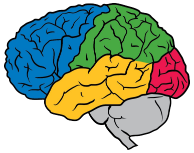
Growing Use of Deep Learning at Google

of directories containing model description files



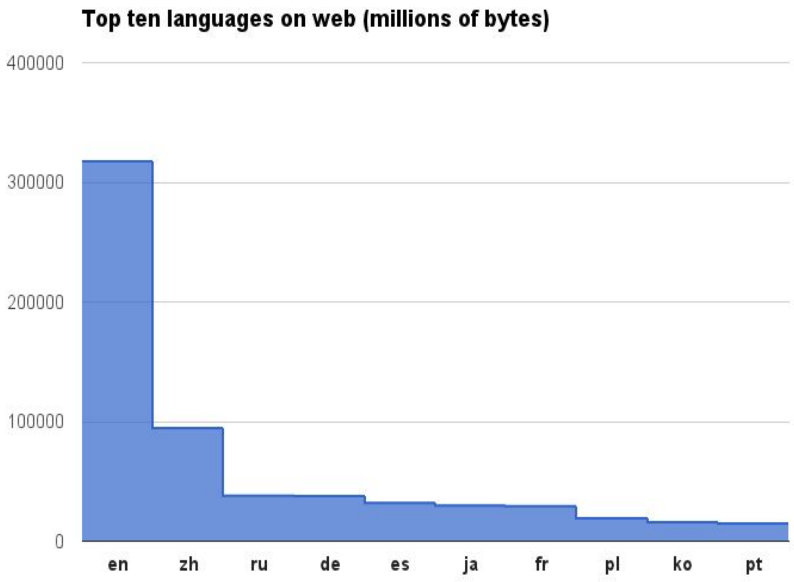
Across many products/areas:

Android
Apps
GMail
Image Understanding
Maps
NLP
Photos
Speech
Translation
many research uses..
YouTube
… many others ...



Why we care about translations

- **50%** of Internet content is in English.
- Only **20%** of the world's population speaks English.



To make the world's information accessible, we need translations

Google Translate, a truly global product...

1B+	Translations every single day, that is 140 Billion Words
1B+	Monthly active users
103	Google Translate Languages cover 99% of online population

Agenda

- Quick History
- From Sequence to Sequence-to-Sequence Models
- BNMT (Brain Neural Machine Translation)
    - Architecture & Training
    - Segmentation Model
    - TPU and Quantization
- Multilingual Models
- What’s next?

Quick Research History

- **Various people at Google tried to improve translation with neural networks**
  - Brain team, Translate team

Quick Research History

- Various people at Google tried to improve translation with neural networks
  - Brain team, Translate team

- **Sequence-To-Sequence models (NIPS 2014)**
  - Based on many earlier approaches to estimate P(Y|X) directly
  - State-of-the-art on WMT En->Fr using custom software, very long training
  - Translation could be learned without explicit alignment!
  - Drawback: all information needs to be carried in internal state
    - Translation breaks down for long sentences!

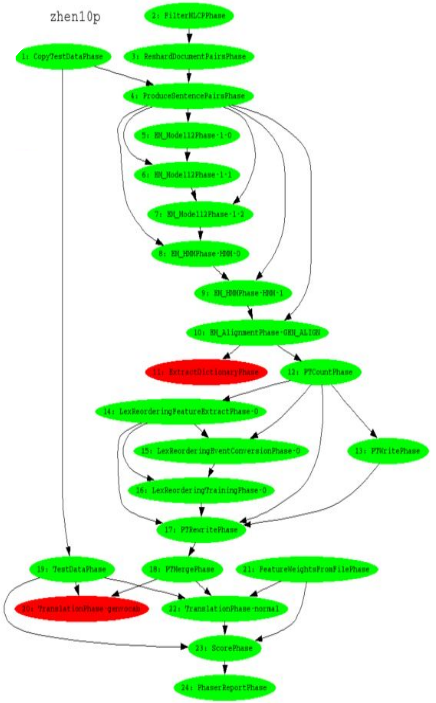
Quick Research History

- Various people at Google tried to improve translation with neural networks
    - Brain team, Translate team

- Sequence-To-Sequence models (NIPS 2014)
    - Based on many earlier approaches to estimate P(Y|X) directly
    - State-of-the-art on WMT En->Fr using custom software, very long training
    - Translation could be learned without explicit alignment!
    - Drawback: all information needs to be carried in internal state
        - Translation breaks down for long sentences!
- **Attention Models (2014)**
    - Removes drawback by giving access to all encoder states
        - Translation quality is now independent of sentence length!

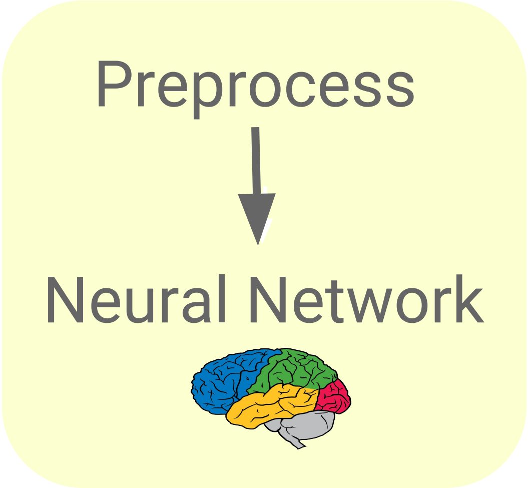
Old: Phrase-based translation

- Lots of individual pieces
- Optimized somewhat independently



New: Neural machine translation

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



Expected time to launch:

3 years

Actual time to launch:

13.5 months

Sept 2015:
Began project using TensorFlow

Feb 2016:
First production data results

Sept 2016:
zh->en launched

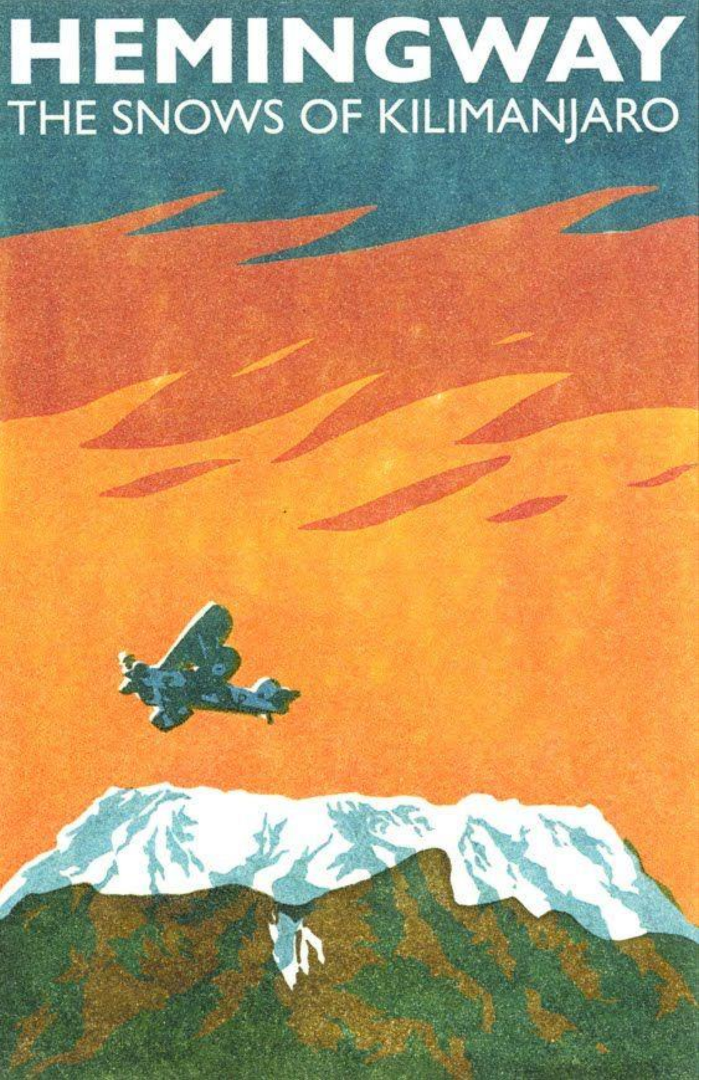
Nov 2016:
8 languages launched
(16 pairs to/from English)

Mar 2017:
7 more launched
(Hindi, Russian, Vietnamese, Thai, Polish, Arabic, Hebrew)

Apr 2017:
26 more launched
(16 European, 8 Indish, Indonesian, Afrikaans)

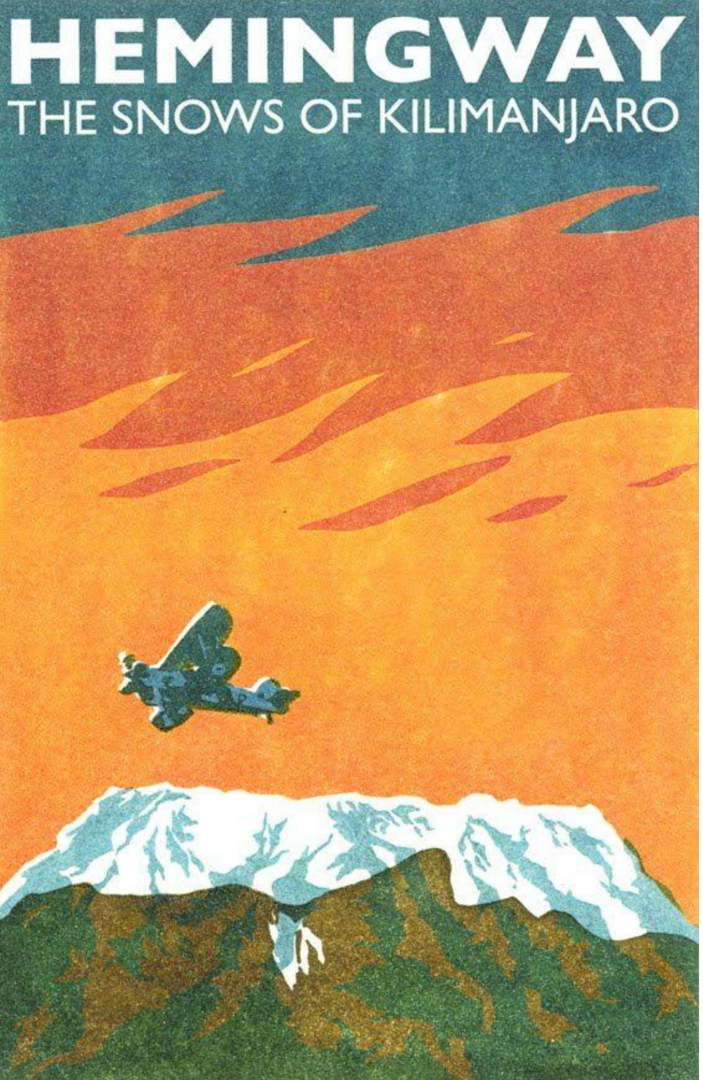
Jun/Aug 2017:
36/20 more launched

97 launched!



Original

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.

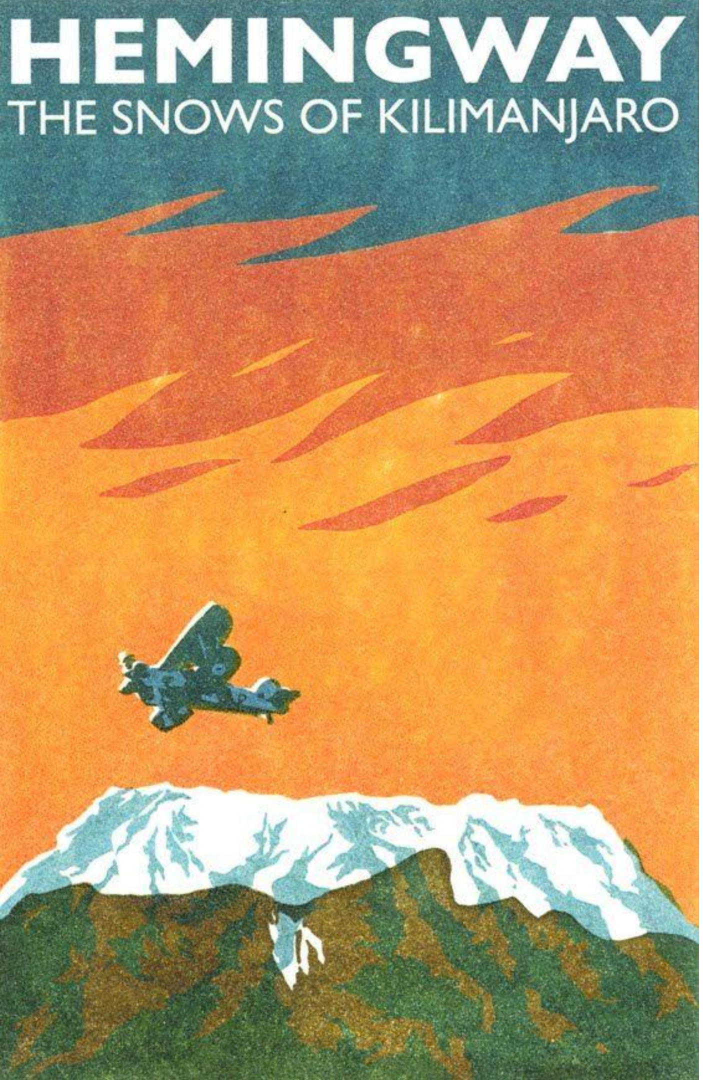


Original

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.



Original

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)

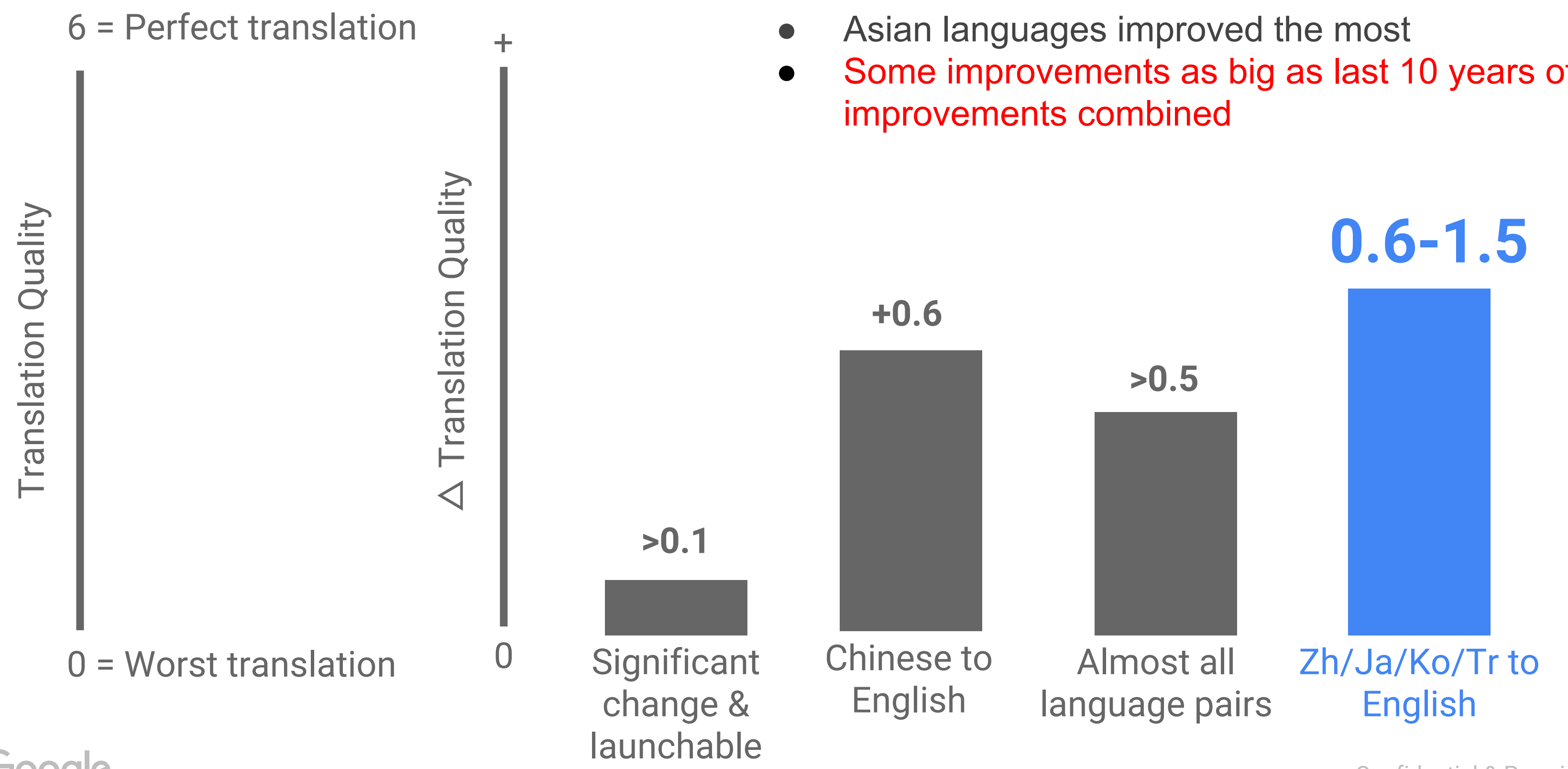
Kilimanjaro is 19,710 feet of the mountain covered with snow, and it is said that the highest mountain in Africa. Top of the west, “Ngaje Ngai” in the Maasai language, has been referred to as the house of God. The top close to the west, there is a dry, frozen carcass of a leopard. Whether the leopard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

Back translation from Japanese (new)

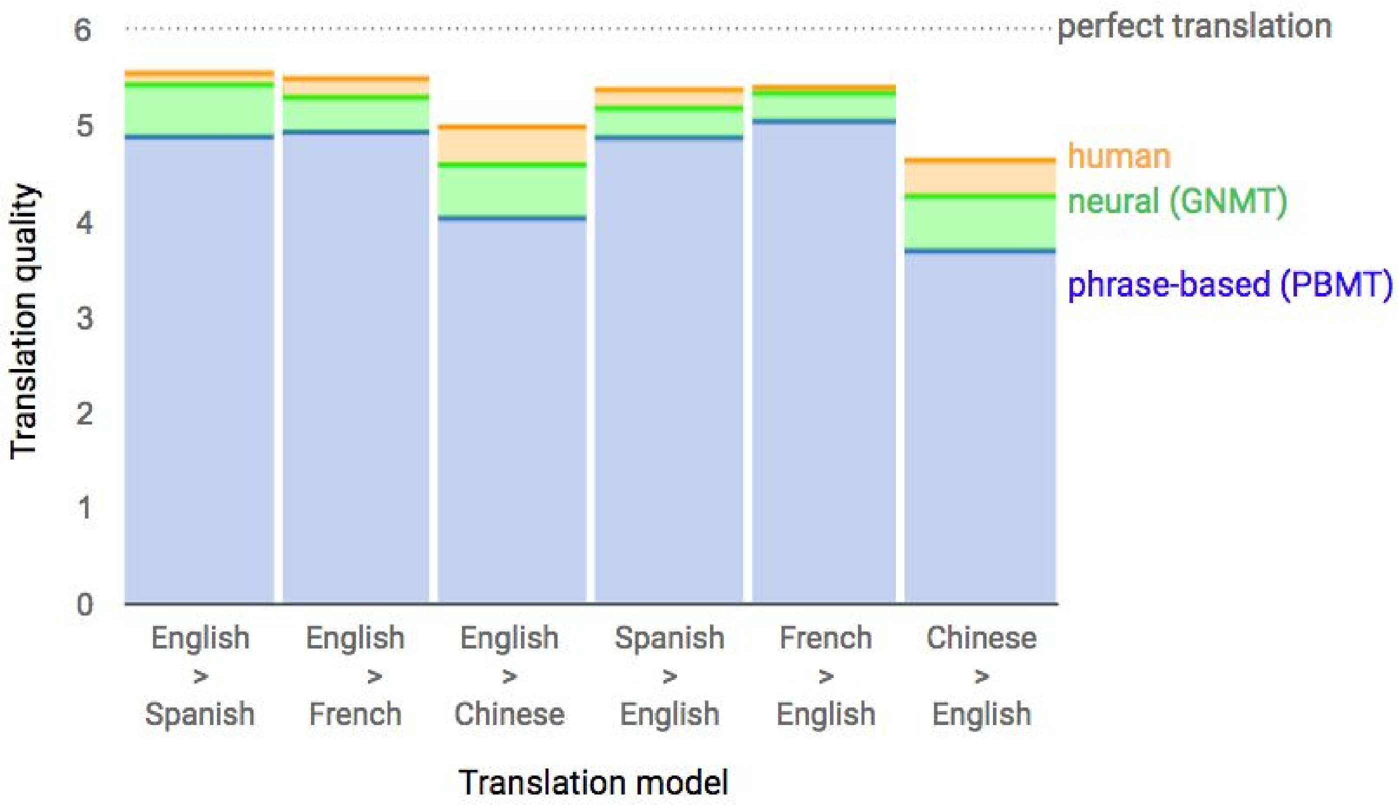
Kilimanjaro is a mountain of 19,710 feet covered with snow, which is said to be the highest mountain in Africa. The summit of the west is called “Ngaje Ngai” God‘s house in Masai language. There is ~~a~~ dried and frozen carcass of a leopard near the summit of the west. No one can explain what ~~the~~ leopard was seeking at that altitude.

Translation Quality

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



Relative Error Reduction



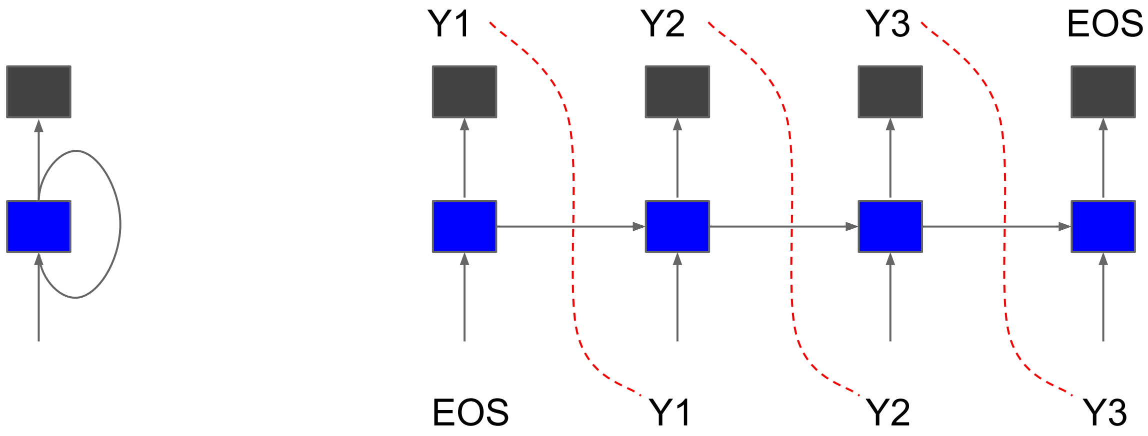
Does quality matter?

+75%

Increase in daily English - Korean translations on Android over the past six months

Neural Recurrent Sequence Models

- Predict next token: P(Y) = P(Y1) * P(Y2|Y1) * P(Y3|Y1,Y2) * ...
  - Language Models, state-of-the-art on public benchmark
    - *Exploring the limits of language modeling*

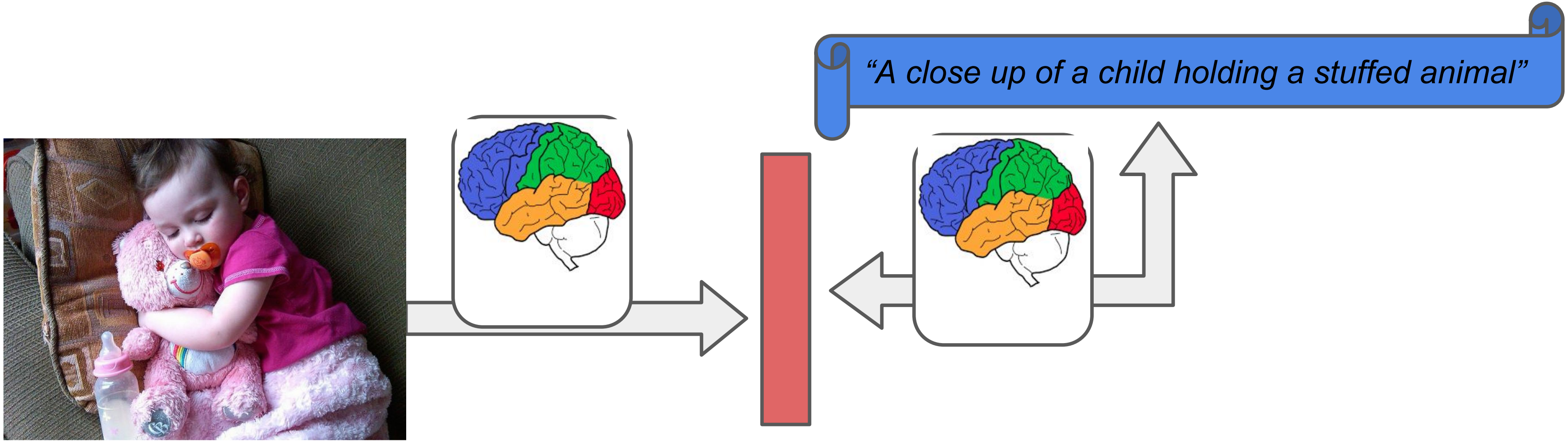


Applications

- Speech Recognition
  - Estimate state posterior probabilities per 10ms frame
- Video Recommendations
  - With hierarchical softmax and MaxEnt model for top 500k YouTube videos

Image Captioning

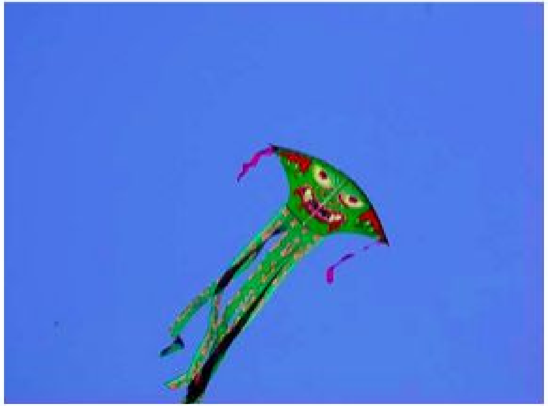
- Combine image classification and sequence model
  - Feed output from image classifier and let it predict text
  - *Show and Tell: A Neural Image Caption Generator*



A man holding a tennis racquet on a tennis court.

Two pizzas sitting on top of a stove top oven

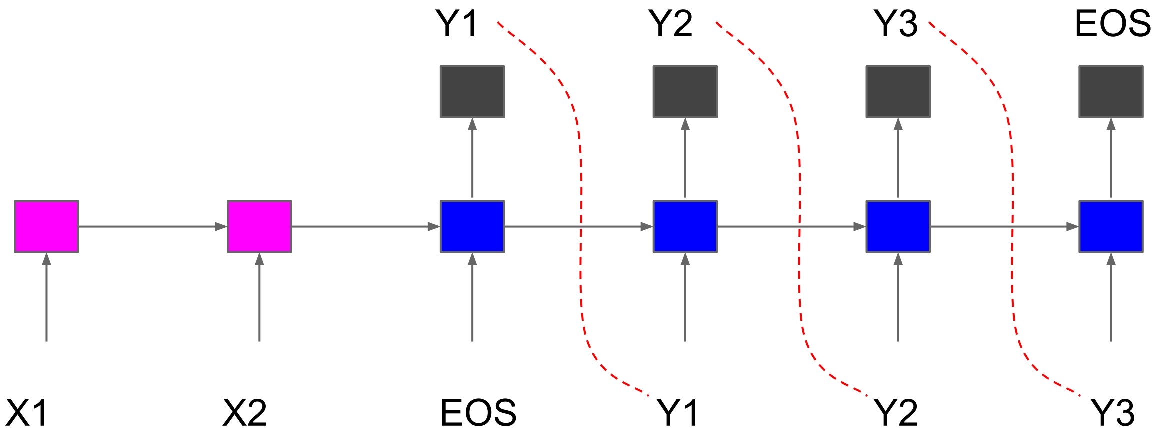
A group of young people playing a game of Frisbee



A man flying through the air while riding a snowboard

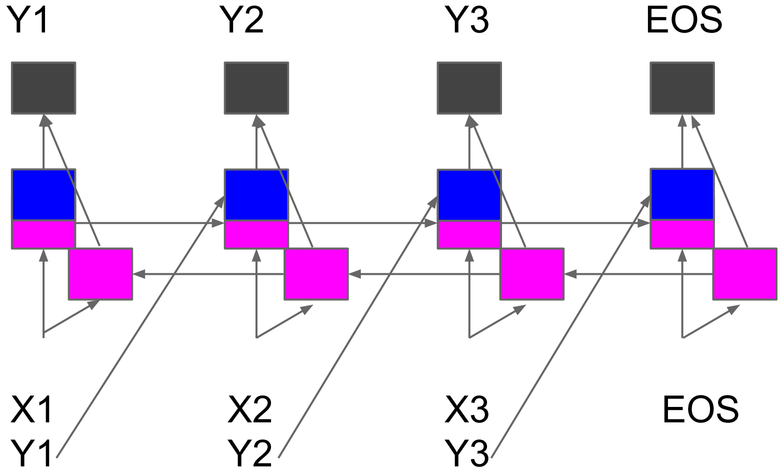
Sequence to Sequence

- Learn to map: X1, X2, EOS -> Y1, Y2, Y3, EOS
- Encoder/Decoder framework (decoder by itself just neural LM)
- Theoretically any sequence length for input/output works

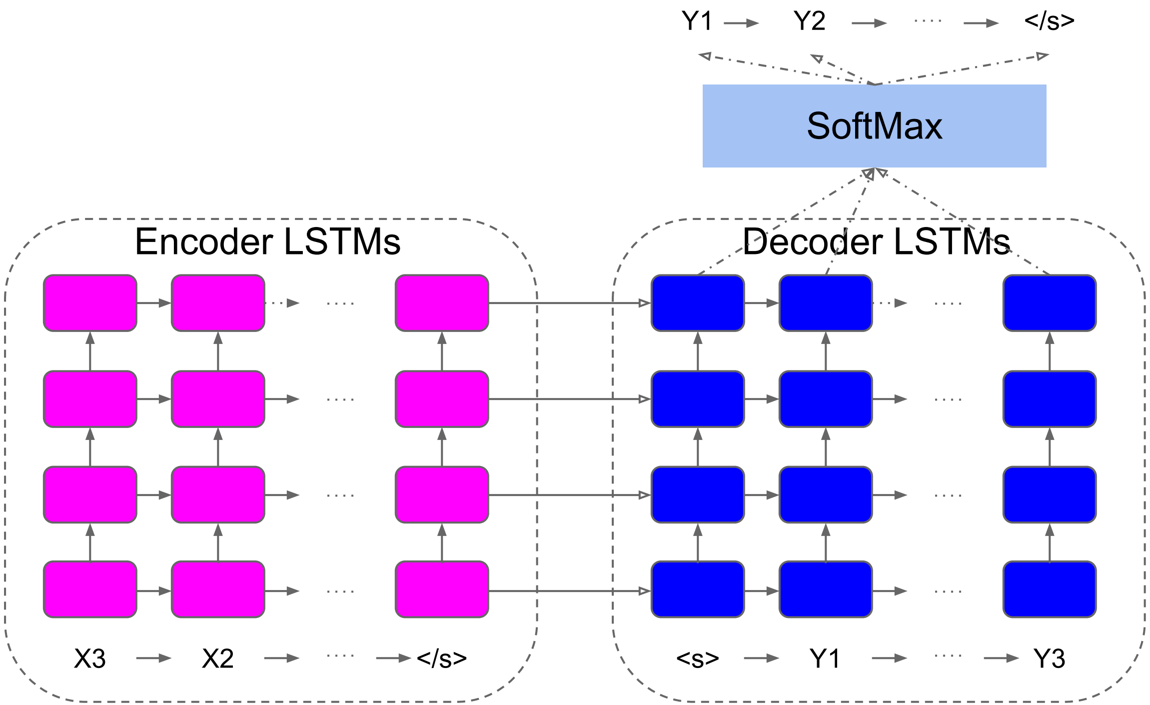


Sequence to Sequence in 1999...

- NN for estimating directly P(Y|X) for ***equal*** length X and Y
- Encoder (BRNN)/Decoder framework but in a single NN
- NIPS 1999
  - *Better Generative Models for Sequential Data Problems: Bidirectional Recurrent Mixture Density Networks*



Deep Sequence to Sequence



Attention Mechanism

- Addresses the information bottleneck problem
    - ***All*** encoder states accessible instead of only final one

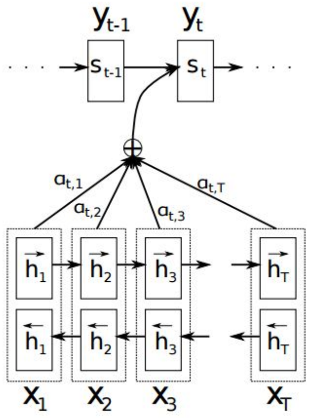
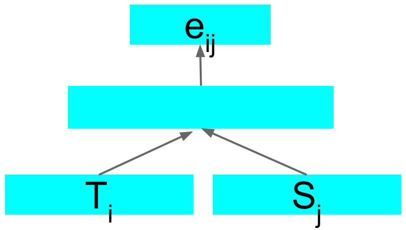
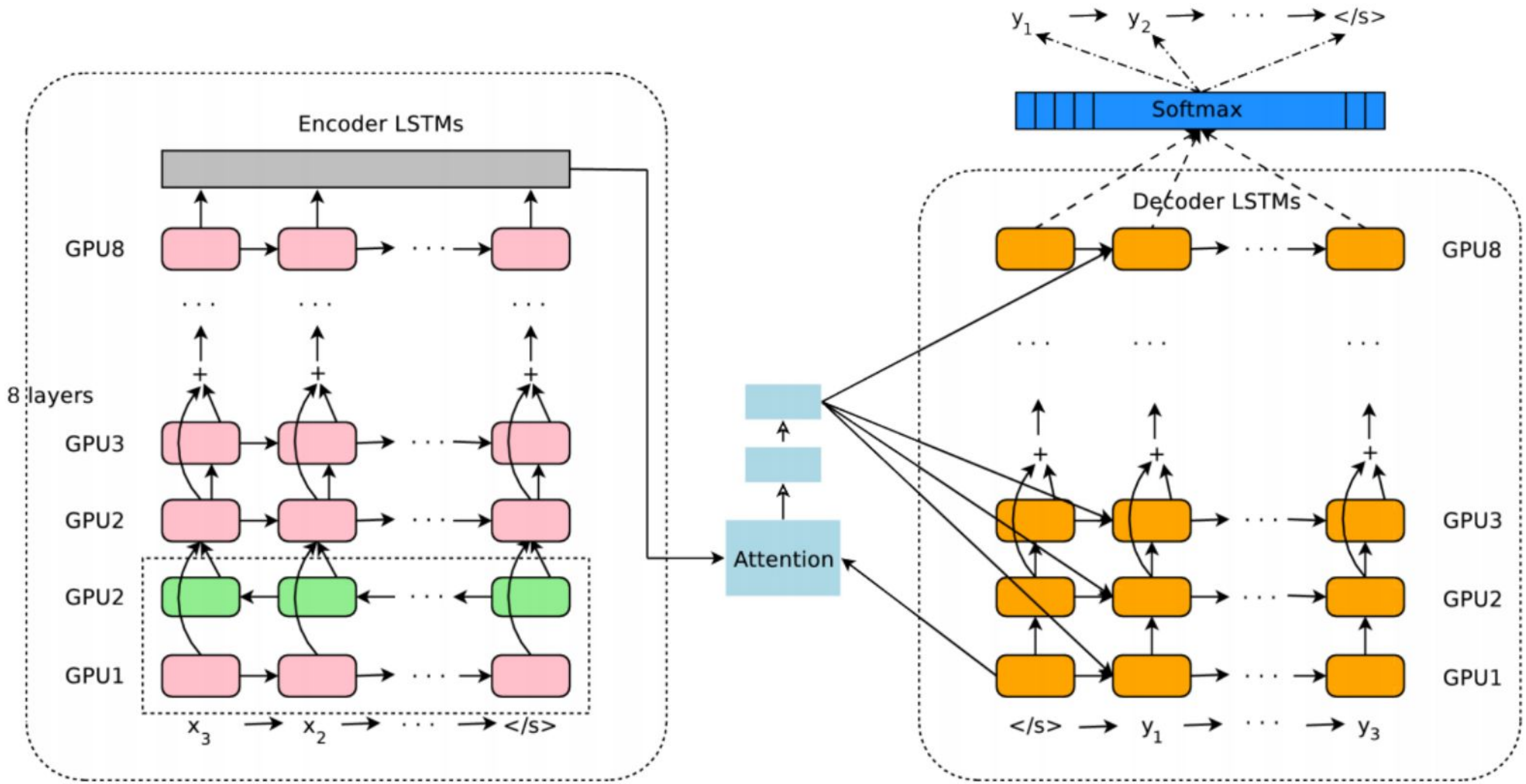


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$



BNMT Model Architecture



Model Training

- **Runs on ~100 GPUs (12 replicas, 8 GPUs each)**
    - Because softmax size only 32k, can be fully calculated (no sampling or HSM)

Model Training

- Runs on ~100 GPUs (12 replicas, 8 GPUs each)
  - Because softmax size only 32k, can be fully calculated (no sampling or HSM)
- **Optimization**
  - Combination of Adam & SGD with delayed exponential decay
  - 128/256 sentence pairs combined into one batch (run in one ‘step’)

Model Training

- Runs on ~100 GPUs (12 replicas, 8 GPUs each)
  - Because softmax size only 32k, can be fully calculated (no sampling or HSM)
- Optimization
  - Combination of Adam & SGD with delayed exponential decay
  - 128/256 sentence pairs combined into one batch (run in one ‘step’)
- **Training time**
  - ~1 week for 2.5M steps = ~300M sentence pairs
  - For example, on English->French we use only 15% of available data!

Wordpiece Model (WPM)

- **Dictionary too big (~100M unique words!)**
  - Cut words into smaller units

Wordpiece Model (WPM)

- Dictionary too big (~100M unique words!)
  - Cut words into smaller units
- **Data-driven bottom-up segmenter (trained once on example data)**
  - Produces predetermined number of units to represent any word possible
  - No UNK (unknown word) problem
  - Frequent words become full units, rare words split up

Wordpiece Model (WPM)

- Dictionary too big (~100M unique words!)
  - Cut words into smaller units
- Data-driven bottom-up segmenter (trained once on example data)
  - Produces predetermined number of units to represent any word possible
  - No UNK (unknown word) problem
  - Frequent words become full units, rare words split up
- **Example**
  - Segmentation
    - add underscore before words, then segment using trained WPM model
      - This is a house -> _Th is _is _a _hou se
  - Desegmentation
    - remove spaces, replace underscore by space
      - _Th is _is _a _hou se -> This is a house

Wordpiece Model (WPM)

- Dictionary too big (~100M unique words!)
  - Cut words into smaller units
- Data-driven bottom-up segmenter (trained once on example data)
  - Produces predetermined number of units to represent any word possible
  - No UNK (unknown word) problem
  - Frequent words become full units, rare words split up
- Example
  - Segmentation
    - add underscore before words, then segment using trained WPM model
      - This is a house -> _Th is _is _a _hou se
  - Desegmentation
    - remove spaces, replace underscore by space
      - _Th is _is _a _hou se -> This is a house
- **Initially developed for speech recognition system (but just like BPE...)**
  - *Japanese and Korean Voice Search*

Wordpiece Model (WPM)

- Particularly important for morphologically rich languages (Ru, De, Ja, Ko, ...)
  - Ru->En: -0.0773 -> +0.462
  - En->Ru: -0.1168 -> +0.259
- Now all languages modeled with WPM (usually 32k)
  - Improves results
  - Lowers latency

Word / Char / Wordpiece / Mixed Word & Char

- Use of WPM improves machine translation measure (BLEU) and lowers latency

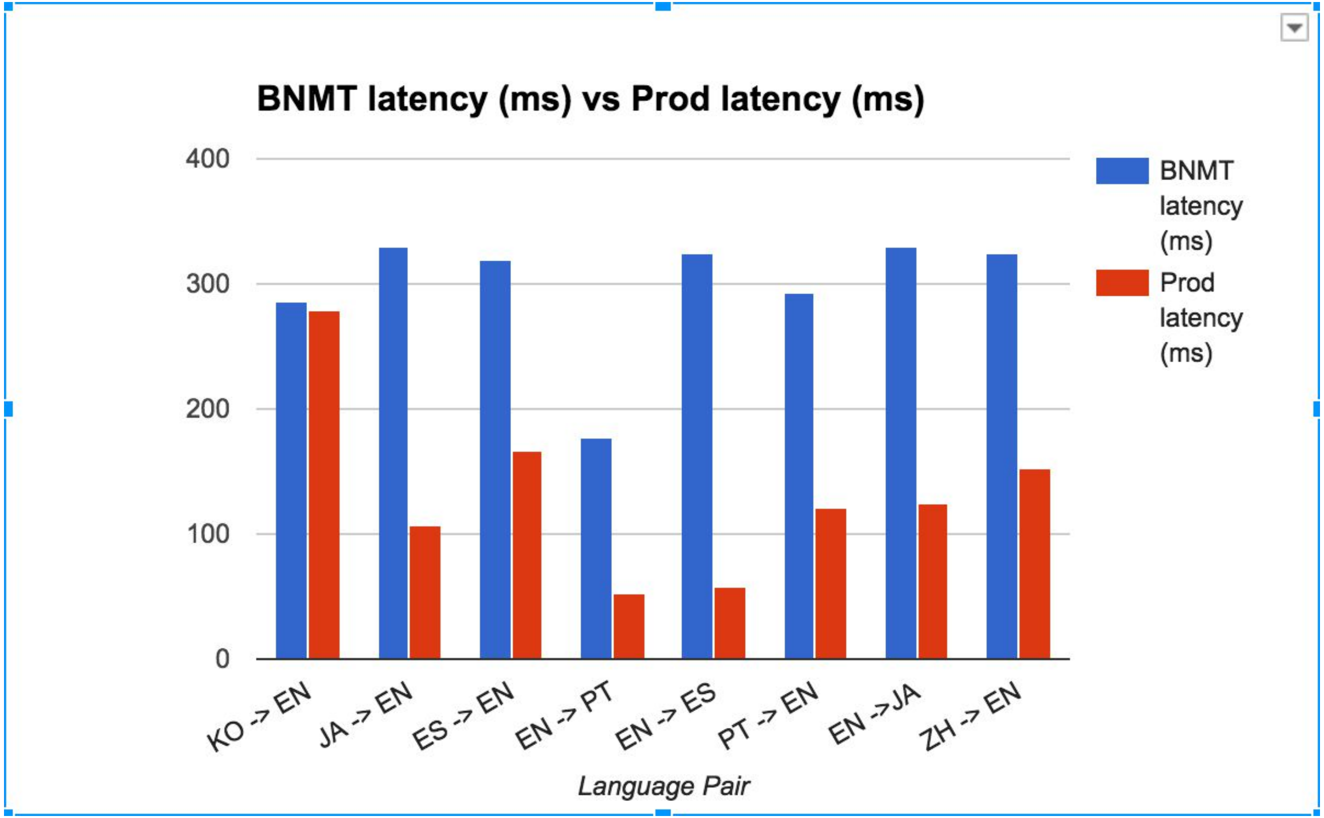
Model (WMT En->Fr)	**BLEU**	**Decoding time/sentence (s)**
Word	37.90	0.2226
Character	38.01	1.0530
WPM-8k	38.27	0.1919
WPM-16k	37.60	**0.1874**
WPM-32k	**38.95**	0.2118
Mixed Word/Character	38.39	0.2774

Speed matters. A lot.

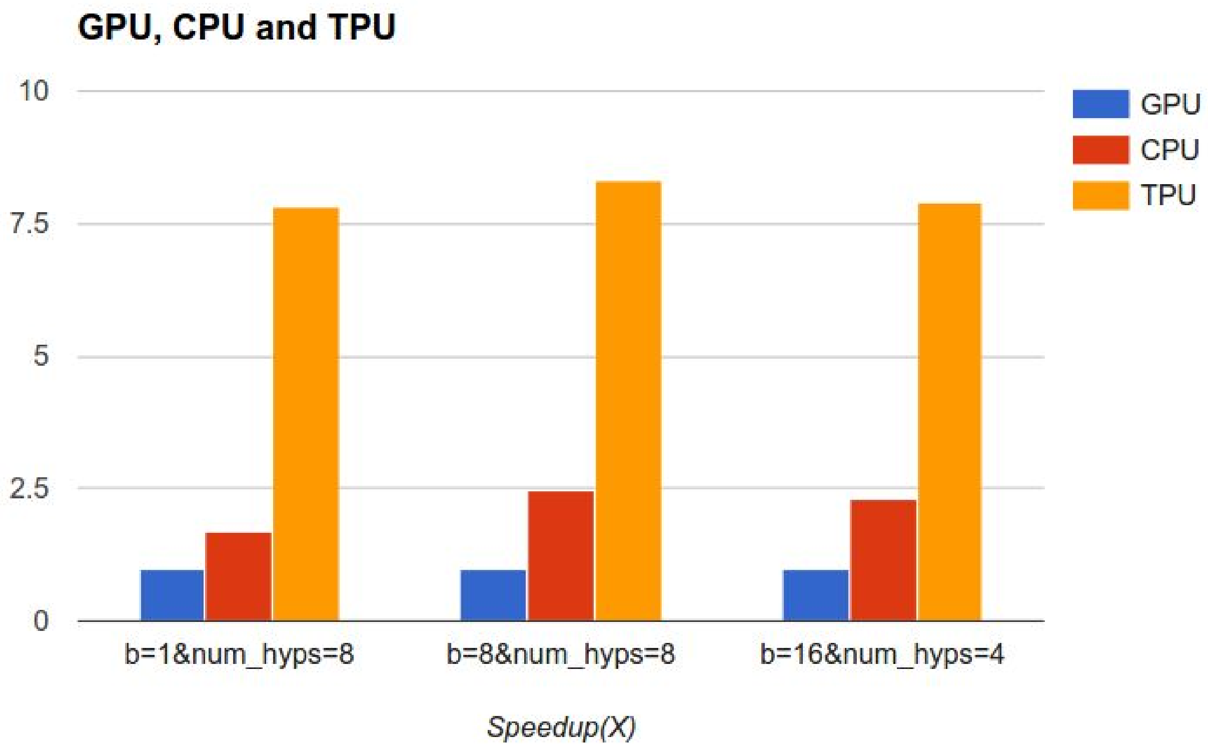
10 seconds/ sentence —— 2 months ——→ 0.2 seconds/ sentence

- Users care about speed
- Better algorithms and hardware (TPUs) made it possible

Latency: BNMT versus PBMT (old system)



Speed-up


GPU, CPU and TPU
10
7.5
5
2.5
0
GPU
CPU
TPU
b=1&num_hyps=8
b=8&num_hyps=8
b=16&num_hyps=4
Speedup(X)

Multilingual Model

- **Model several language pairs in single model**
  - We ran first experiments in 2/2016, surprisingly this worked

Multilingual Model

- Model several language pairs in single model
  - We ran first experiments in 2/2016, surprisingly this worked!
- **Prepend source with additional token to indicate target language**
  - Translate to Spanish:
    - <2es> How are you </s> -> Cómo estás </s>
  - Translate to English:
    - <2en> Como estás </s> -> How are you </s>

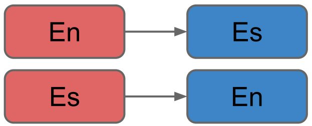
Multilingual Model

- Model several language pairs in single model
  - We ran first experiments in 2/2016, surprisingly this worked!
- Prepend source with additional token to indicate target language
  - Translate to Spanish:
    - <2es> How are you </s> -> Cómo estás </s>
  - Translate to English:
    - <2en> Cómo estás </s> -> How are you </s>

- **No other changes to model architecture!**
  - Extremely simple and effective
  - Usually with shared WPM for source/target

Multilingual Model and Zero-Shot Translation

1.



Single	Multi
34.5	**35.1**
38.0	37.3

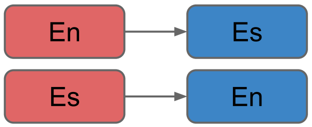
Translation:

<2es> How are you </s> Cómo estás </s>

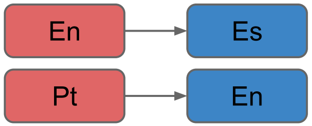
<2en> Cómo estás </s> How are you </s>

Multilingual Model and Zero-Shot Translation

1.



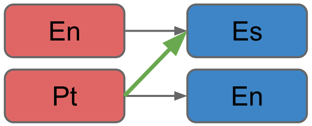
2.



Single	**Multi**
34.5	**35.1**
38.0	37.3
34.5	**35.0**
44.5	43.7

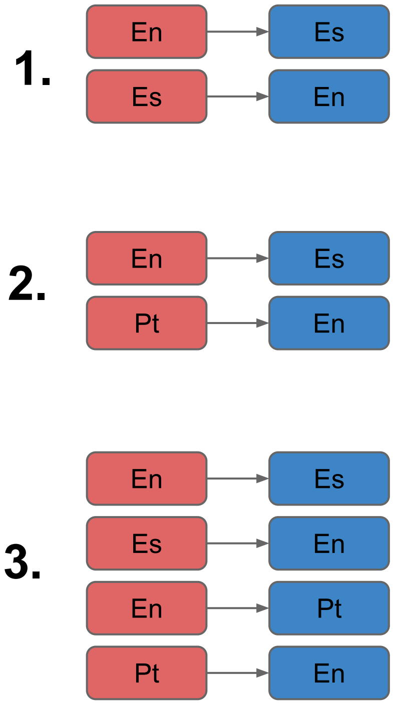
Translation:
<2es> How are you </s> Cómo estás </s>
<2en> Cómo estás </s> How are you </s>

Zero-shot (pt->es):
<2es> Como você está </s> Cómo estás </s>



23.0 BLEU

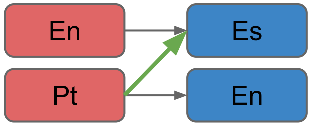
Multilingual Model and Zero-Shot Translation



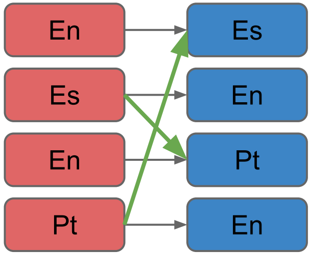
Single	Multi
34.5	**35.1**
38.0	37.3
34.5	**35.0**
44.5	43.7
34.5	**34.9**
38.0	37.2
37.1	**37.8**
44.5	43.7

Translation:
<2es> How are you </s> Cómo estás </s>
<2en> Cómo estás </s> How are you </s>

Zero-shot (pt->es):
<2es> Como você está </s> Cómo estás </s>



23.0 BLEU



24.0 BLEU

Mixing Languages on Source Side

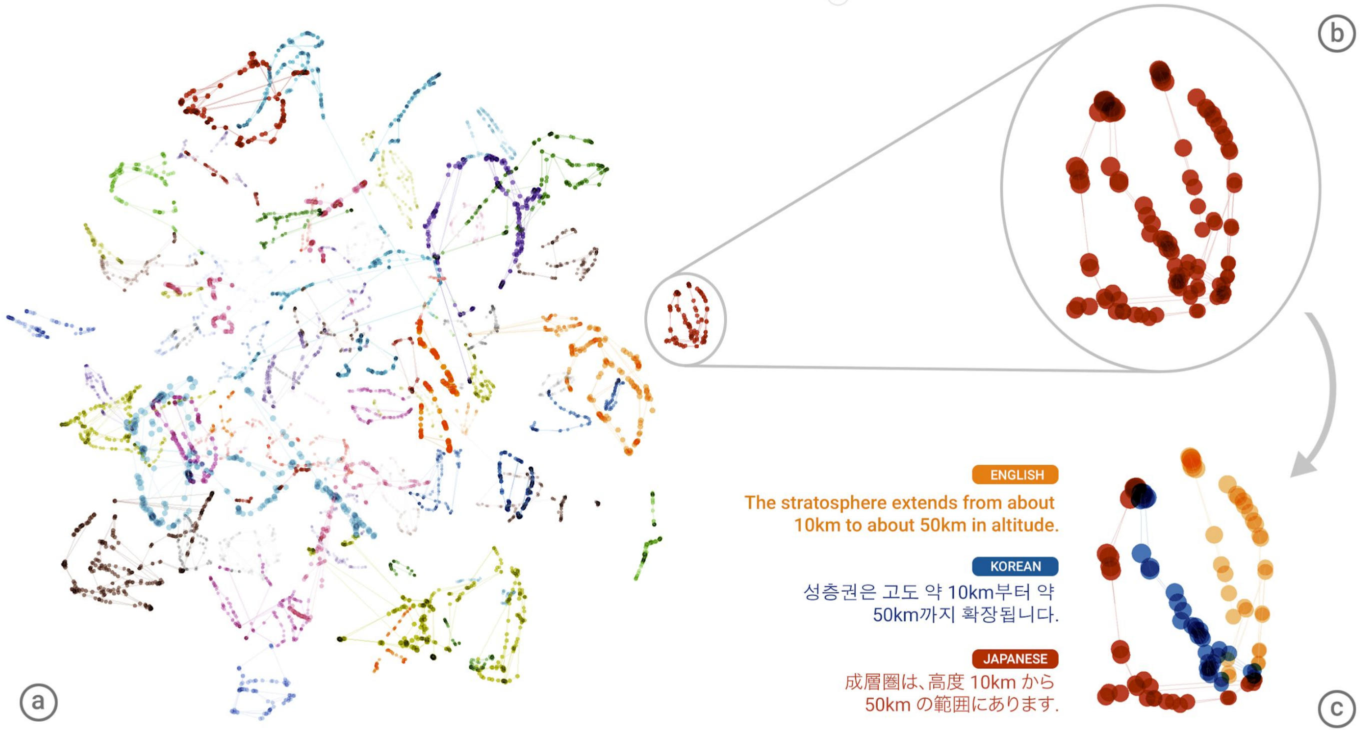
- Code-switching in Japanese/Korean->English model
  - Japanese
    - 私は東京大学の学生です。→ I am a student at Tokyo University.
  - Korean
    - 나는 도쿄 대학의 학생입니다. → I am a student at Tokyo University.
  - Mixed Japanese/Korean
    - 私は東京大学 학생입니다. → I am a student of Tokyo University.

Weighted Target Language Selection

- Linear interpolation of tokens <2ja> and <2ko> (“Japarean” ;-)
  - Model: English->Japanese/Korean
- English: “I must be getting somewhere near the centre of the earth.”
  - w_{ko} = 0.00: 私は地球の中心の近くにどこかに行っているに違いない。
  - w_{ko} = 0.40: 私は地球の中心近くのどこかに着いているに違いない。
  - w_{ko} = 0.56: 私は地球の中心の近くのどこかになっているに違いない。
  - w_{ko} = 0.58: 私は지구の中心의가까이에어딘가에도착하고있어야한다。
  - w_{ko} = 0.60: 나는지구의센터의가까이에어딘가에도착하고있어야한다。
  - w_{ko} = 0.70: 나는지구의중심근처어딘가에도착해야합니다。
  - w_{ko} = 0.90: 나는어딘가지구의중심근처에도착해야합니다。
  - w_{ko} = 1.00: 나는어딘가지구의중심근처에도착해야합니다。
- Other examples go through a third language in the middle!

Interlingua?

Sentences with same meaning mapped to similar regions regardless of language!



Challenges

- **Early cutoff**
  - Cuts off or drops some words in source sentence

Challenges

- Early cutoff
  - Cuts off or drops some words in source sentence
- **Broken dates/numbers**
  - 5 days ago -> 6일 전
  - (but, on average BNMT significantly better than PBMT on number expressions!)

Challenges

- Early cutoff
  - Cuts off or drops some words in source sentence
- Broken dates/numbers
  - 5 days ago -> 6일 전
  - (but, on average BNMT significantly better than PBMT on number expressions!)
- **Short rare queries**
  - “deoxyribonucleic acid” in Japanese ?
  - Should be easy but isn’t yet

Challenges

- Early cutoff
  - Cuts off or drops some words in source sentence
- Broken dates/numbers
  - 5 days ago -> 6일 전
  - (but, on average BNMT significantly better than PBMT on number expressions!)
- Short rare queries
  - “deoxyribonucleic acid” in Japanese ?
  - Should be easy but isn’t yet
- **Transliteration of names**
  - Eichelbergertown -> アイセルベルクタウン

Challenges

- Early cutoff
  - Cuts off or drops some words in source sentence
- Broken dates/numbers
  - 5 days ago -> 6일 전
  - (but, on average BNMT significantly better than PBMT on number expressions!)
- Short rare queries
  - “deoxyribonucleic acid” in Japanese ?
  - Should be easy but isn’t yet
- Transliteration of names
  - Eichelbergertown -> アイセルベルクタウン
- **Junk**
  - xxx -> 牛津词典 (Oxford dictionary)
  - The cat is a good computer. -> 的英语翻译 (of the English language?)
  - Many sentences containing news started with “Reuters”

Open Research Problems

- **Use of context**
  - Full document translation, streaming translation
  - Use other modalities & features

Open Research Problems

- Use of context
  - Full document translation, streaming translation
  - Use other modalities & features
- **Better automatic measures & objective functions**
  - Current BLEU score weighs all words the same regardless of meaning
    - ‘president’ mostly more important than ‘the’
  - Discriminative training
    - Training with Maximum Likelihood produces mismatched training/test procedure!
      - No decoding errors for maximum-likelihood training
    - RL (and similar) already running but no significant enough gains yet
      - Humans see no difference in the results

Open Research Problems

- Use of context
  - Full document translation, streaming translation
  - Use other modalities & features
- Better automatic measures & objective functions
  - Current BLEU score weighs all words the same regardless of meaning
    - ‘president’ mostly more important than ‘the’
  - Discriminative training
    - Training with Maximum Likelihood produces mismatched training/test procedure!
      - No decoding errors for maximum-likelihood training
    - RL (and similar) already running but no significant enough gains yet
      - Humans see no difference in the results

- **Lots of improvements are boring to do!**
  - Because they are incremental (but still have to be done)
  - Data cleaning, new test sets etc.

What’s next from research?

- Convolutional sequence-to-sequence models
    - No recurrency, just windows over input with shared parameters
    - Encoder can be computed in parallel => faster
- Attention only sequence-to-sequence models
    - No recurrency, no convolution, just attention => even simpler!
    - Basic idea: Attention per layer
    - Paper (now on arXiv)
        - *Attention is all you need*

BNMT for other projects

Other projects using same codebase for completely different problems (in search, Google Assistant, …)

- Question/answering system (chat bots)
- Summarization
- Dialog modeling
- Generate question from query
- …

Resources

- TensorFlow (www.tensorflow.org)
    - Code/Bugs on GitHub
    - Help on StackOverflow
    - Discussion on mailing list
- All information about BNMT is in these papers & blog posts
    - *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*
    - *Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation*
- NYT article describes some of the development
    - *The Great AI Awakening*
- Internship & Residency
    - 3 months internships possible
    - 1-year residency program g.co/brainresidency

Questions?

Thank you!

schuster@google.com

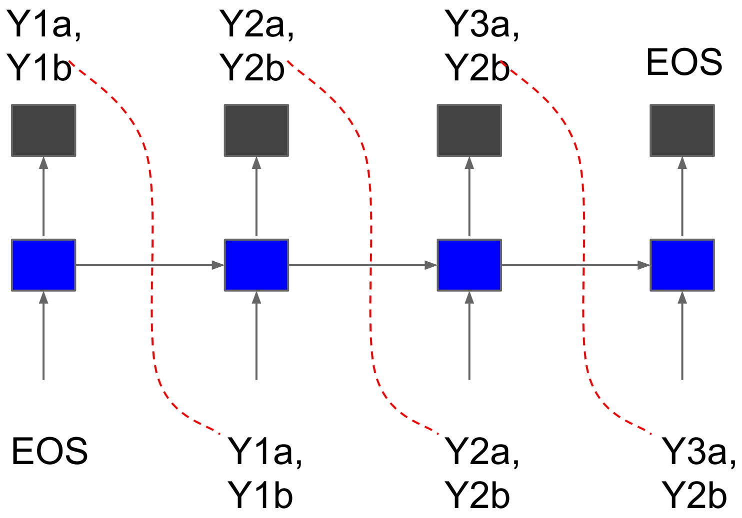
g.co/brain

Decoding Sequence Models

- Find the N-best highest probability output sequences
  - Take K-best Y1 and feed them one-by-one, generating K hypotheses
  - Take K-best Y2 for each of the hyps, generating K^2 new hyps (tree) etc.
  - At each step, cut hyps to N-best (or by score) until at end

Example N=2, K=2

1. Y1a Y2a …
2. Y1a Y2b …
3. Y1b Y2a …
4. Y1b Y2b ...



Sampling from Sequence Models

- Generate samples of sequences
  a. Generate probability distribution P(Y1)
  b. Sample from P(Y1) according to its probabilities
  c. Feed in found sample as input, goto a)