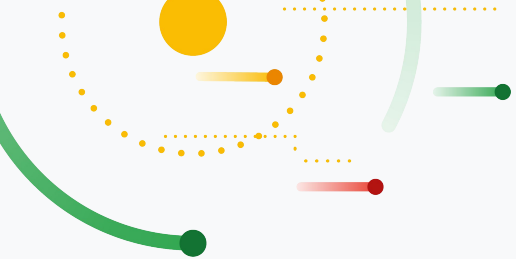
Google AI

The Many Machine Learning Ways of Google Translate

Keith Stevens @ Google Translate
kstevens@google.com



Google AI

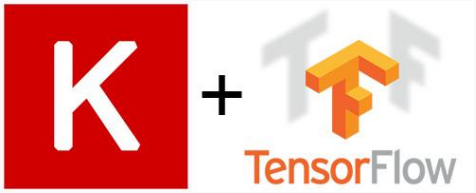
Machine Learning at Google

Growing Ecosystem

- Core Framework for building models

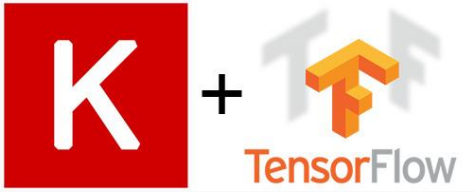
Growing Ecosystem

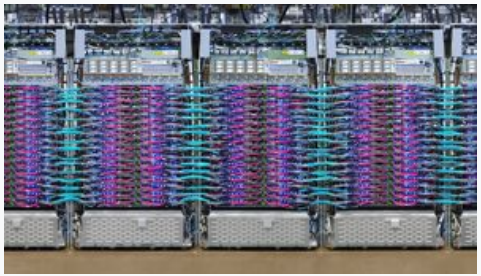
- Core Framework for building models
- Improving usability with Keras and other APIs



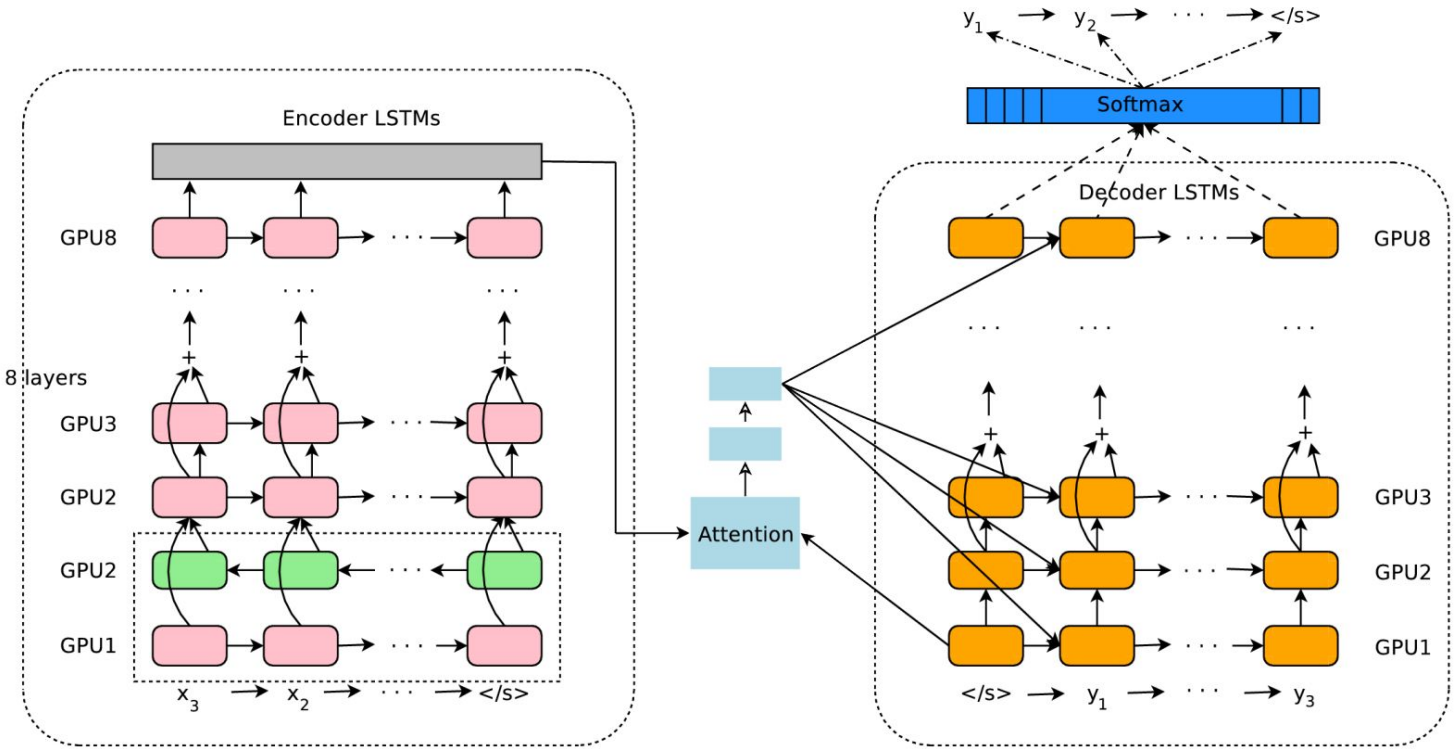
Growing Ecosystem

- Core Framework for building models
- Improving usability with Keras and other APIs
- Enabling scale with cloud Tensor Processing Units





Translate's Sequence to Sequence model



Benchmarking Machine Translation progress in 2 years

Sur un circuit quipromet un sprint massif à l'arrivée,
Bourreau a choisi d'emmener deux leaders avec
autour d'eux leurs habituels équipiers.

Query

Benchmarking Machine Translation progress in 2 years

Sur un circuit quipromet un sprint massif à l'arrivée,
Bourreau a choisi d'emmener deux leaders avec
autour d'eux leurs habituels équipiers.

Query

On a circuit that promises a sprint to the finish ,
Executioner chose to bring both leaders with
their teammates around them usual .

Phrase Based Translation

Benchmarking Machine Translation progress in 2 years

Sur un circuit quipromet un sprint massif à l'arrivée,
Bourreau a choisi d'emmener deux leaders avec
autour d'eux leurs habituels équipiers.

Query

On a circuit that promises a sprint to the finish , Executioner chose to bring both leaders with their teammates around them usual .

Phrase Based Translation

On a circuit that promises a massive sprint at the finish, Bourreau has chosen to take two leaders with their usual crew around them.

Neural Machine Translation (~2017)

Benchmarking Machine Translation progress in 2 years

Sur un circuit quipromet un sprint massif à l'arrivée, Bourreau a choisi d'emmener deux leaders avec autour d'eux leurs habituels équipiers.

Query

On a circuit that promises a sprint to the finish , Executioner chose to bring both leaders with their teammates around them usual .

Phrase Based Translation

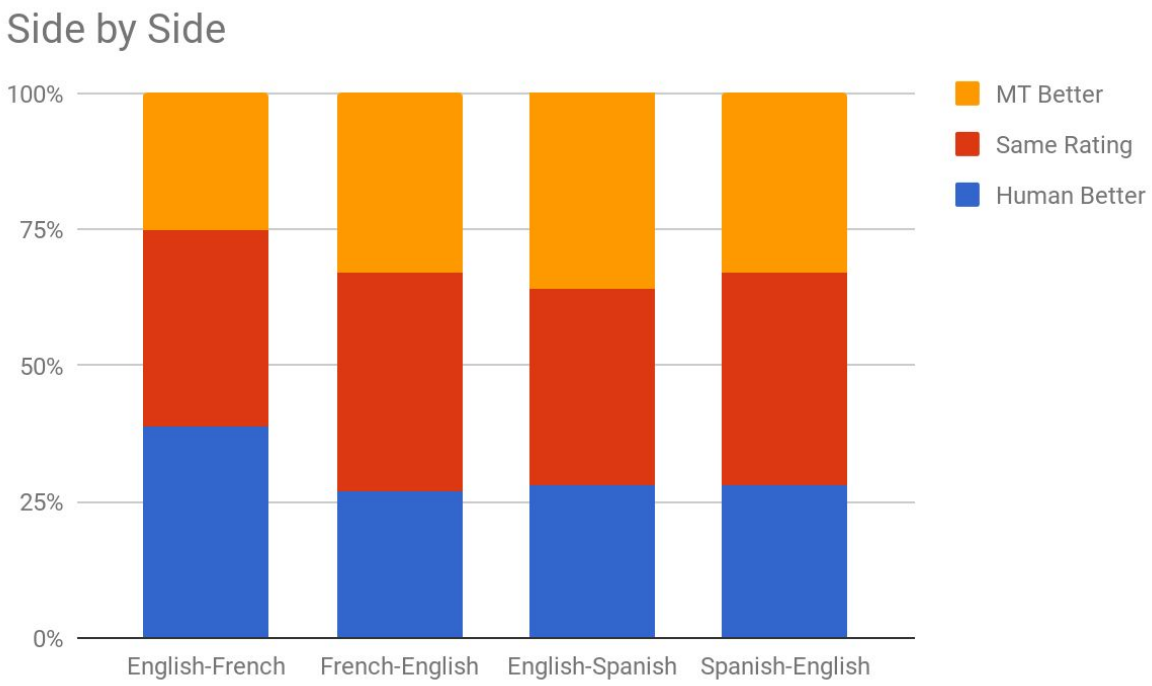
On a circuit that promises a massive sprint at the finish, Bourreau has chosen to take two leaders with their usual crew around them.

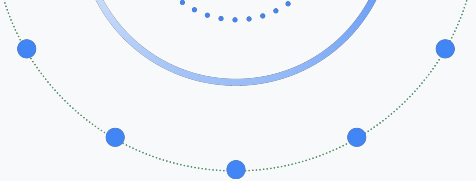
Neural Machine Translation (~2017)

On a circuit that starts a massive sprint at the finish, Bourreau has chosen to bring two leaders with around them their usual teammates.

Neural Machine Translation (Now)

Positive Metrics don't always reveal success



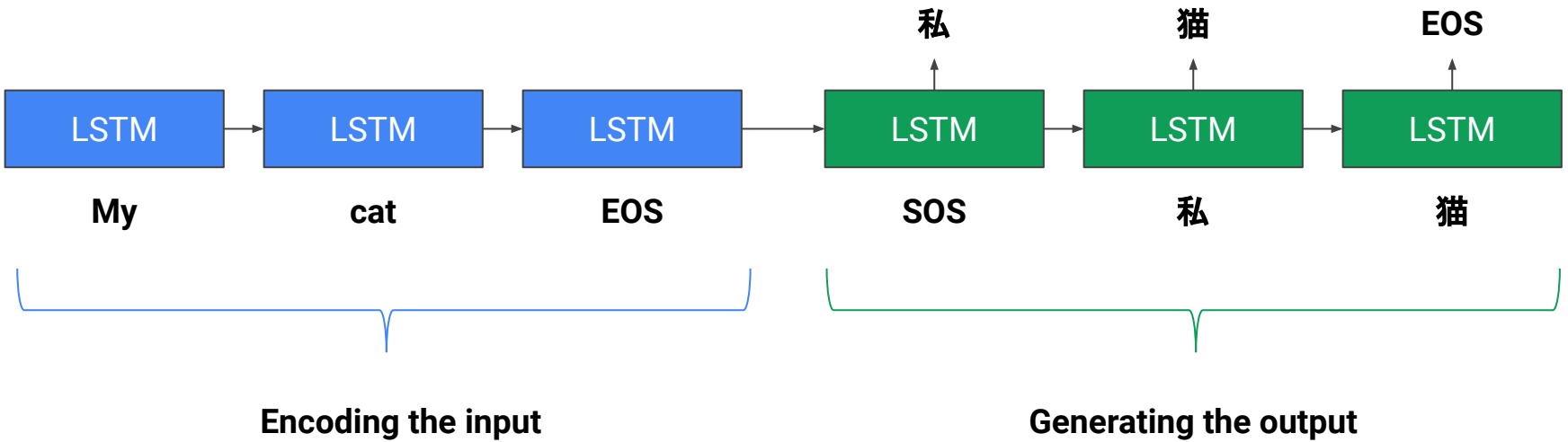


Google AI

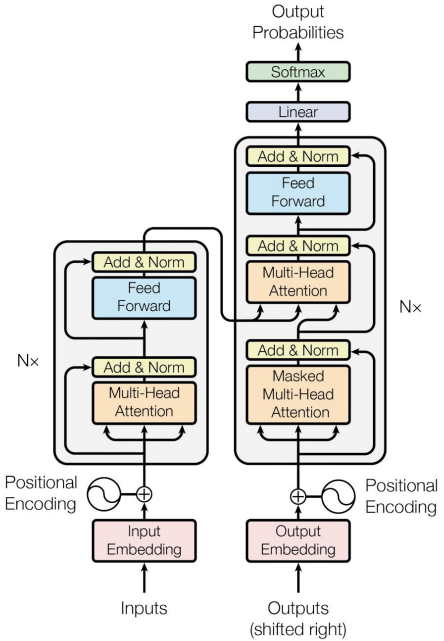
The Well Known Big Changes

High level Overview

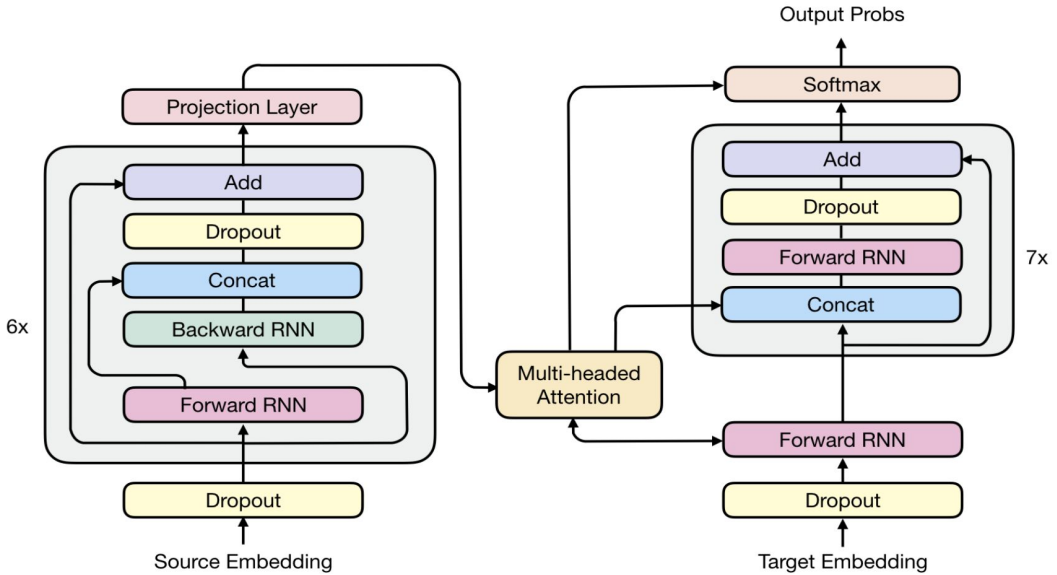
It all started with LSTMs



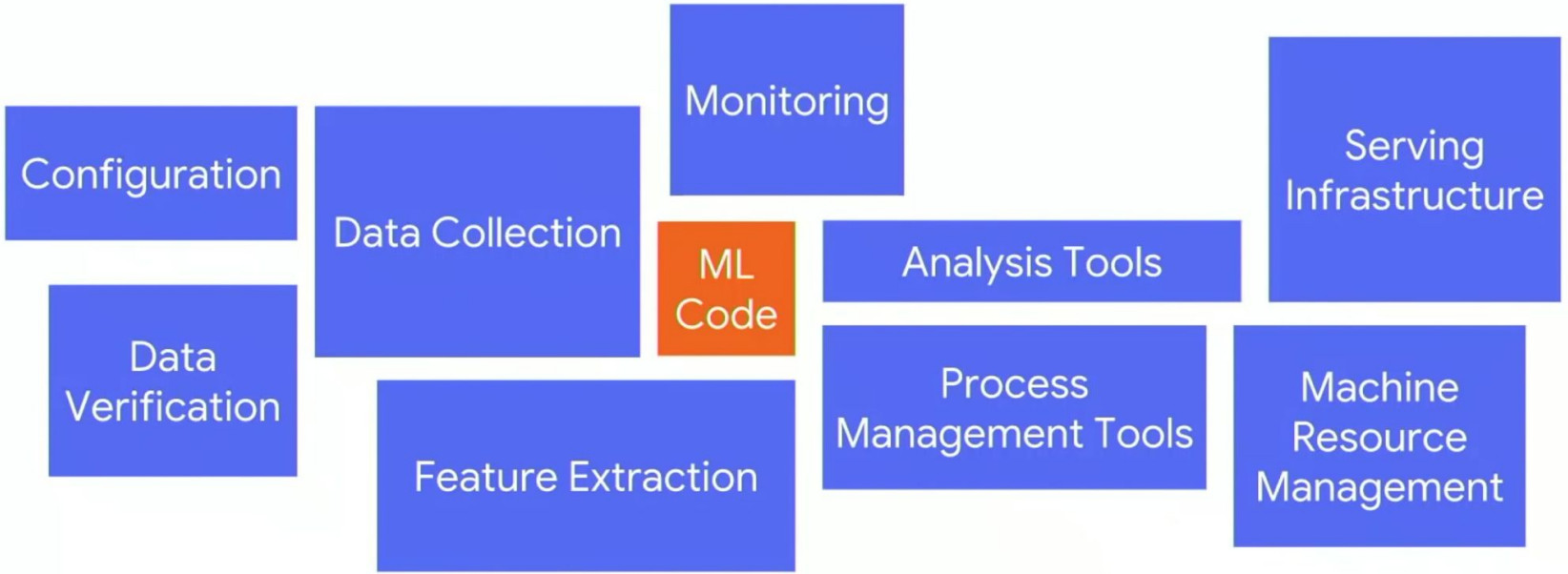
Then we got way more complex

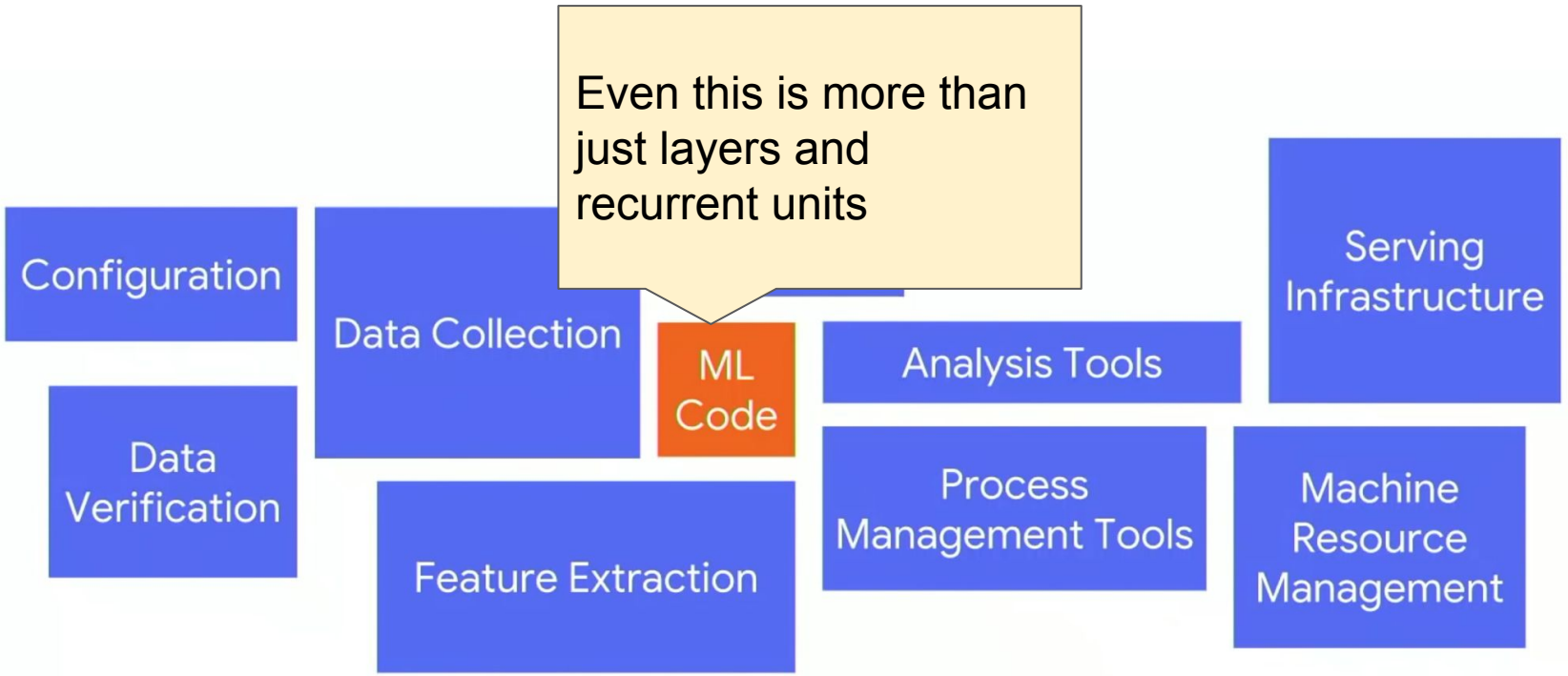


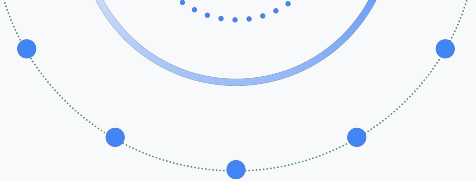
Transformers



Hybrid


Configuration
Data Collection
Monitoring
Serving Infrastructure
Data Verification
ML Code
Analysis Tools
Process Management Tools
Machine Resource Management
Feature Extraction


Even this is more than just layers and recurrent units
Configuration
Data Collection
ML Code
Analysis Tools
Serving Infrastructure
Data Verification
Feature Extraction
Process Management Tools
Machine Resource Management



Machine Learning beyond Architecture Search

Overview of recent work (mostly using Lingvo)

A sample of unsolved problems

Pushing the limits of model size

Working with non-ideal data

Translating on the fly

Using models to find data

Pushing the limits of model size

Open Question: Can we train one model for all languages?

Papers:

Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges

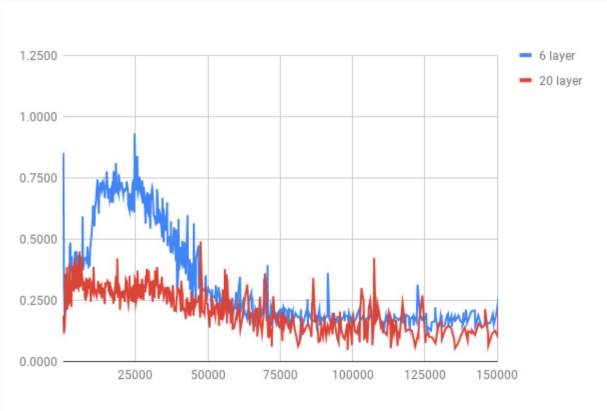
Training Deeper Neural Machine Translation Models with Transparent Attention

Pushing the limits of model size

Open Question: Can we train one model for all languages?

Problems:

- 20 layers breaks Transformers



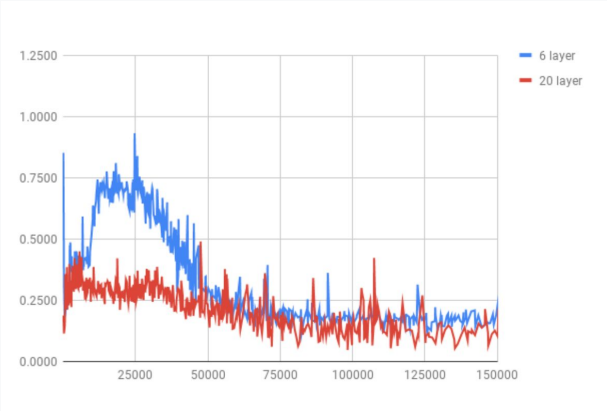
Transformer training breaks down with 20 layers

Pushing the limits of model size

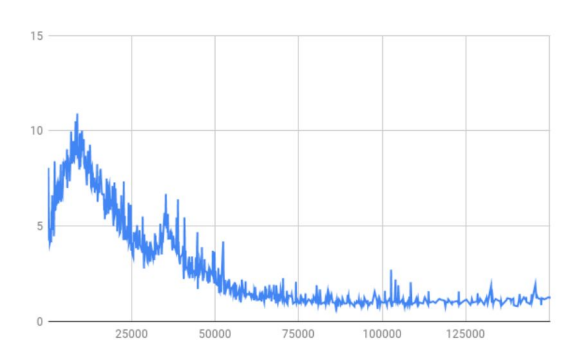
Open Question: Can we train one model for all languages?

Problems:

- 20 layers breaks Transformers



Transformer training breaks down with 20 layers



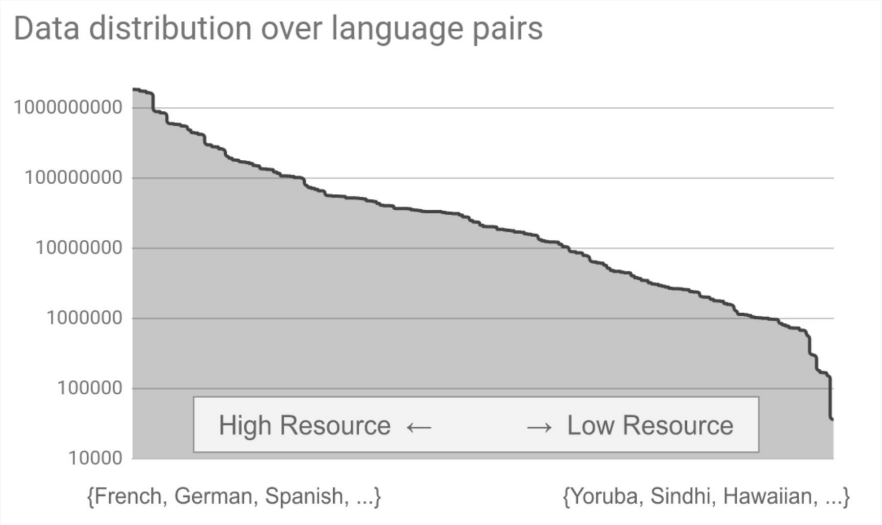
Transformer training stays stable using Transparent Attention

Pushing the limits of model size

Open Question: Can we train one model for all languages?

Problems:

- 20 layers breaks Transformers
- Uneven datasets -> Must Balance



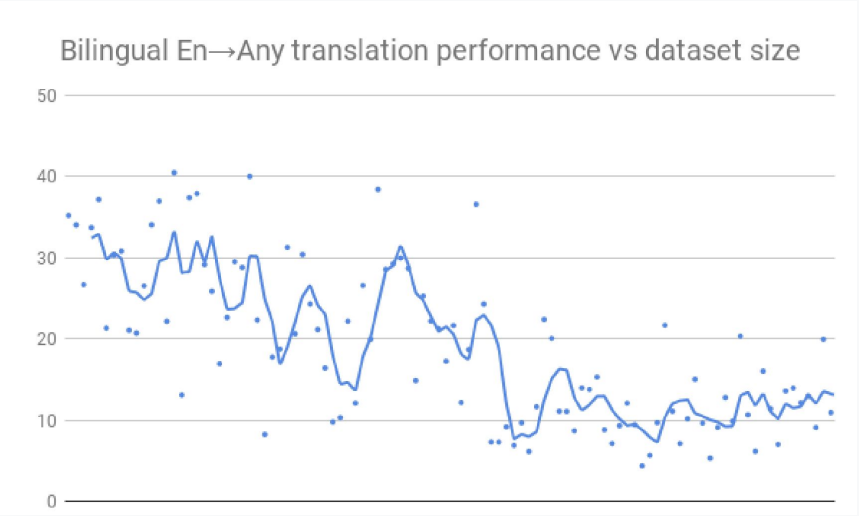
Nearly 5 orders of magnitude difference in size

Pushing the limits of model size

Open Question: Can we train one model for all languages?

Problems:

- 20 layers breaks Transformers
- Uneven datasets -> Must Balance
- Mixed quality across languages



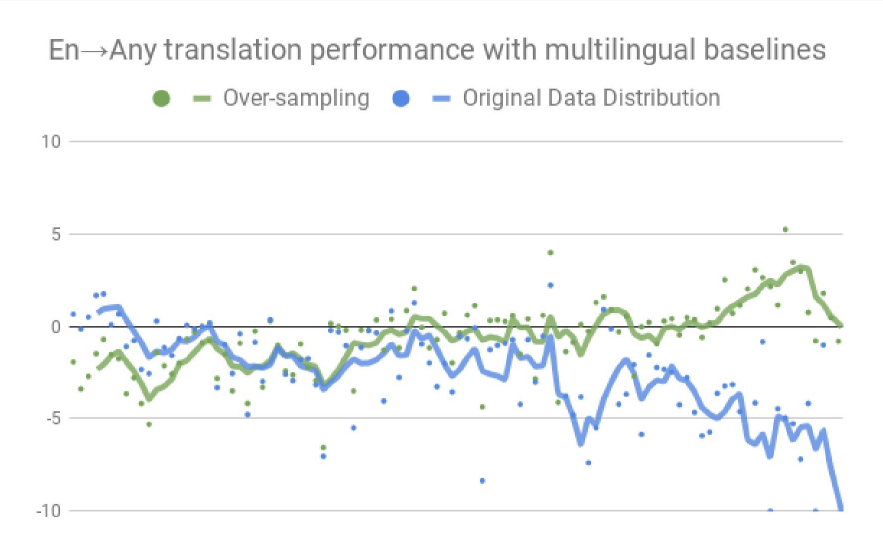
Lack of high quality data for many languages

Pushing the limits of model size

Open Question: Can we train one model for all languages?

Lessons Learned:

- Oversampling works & doesn't work



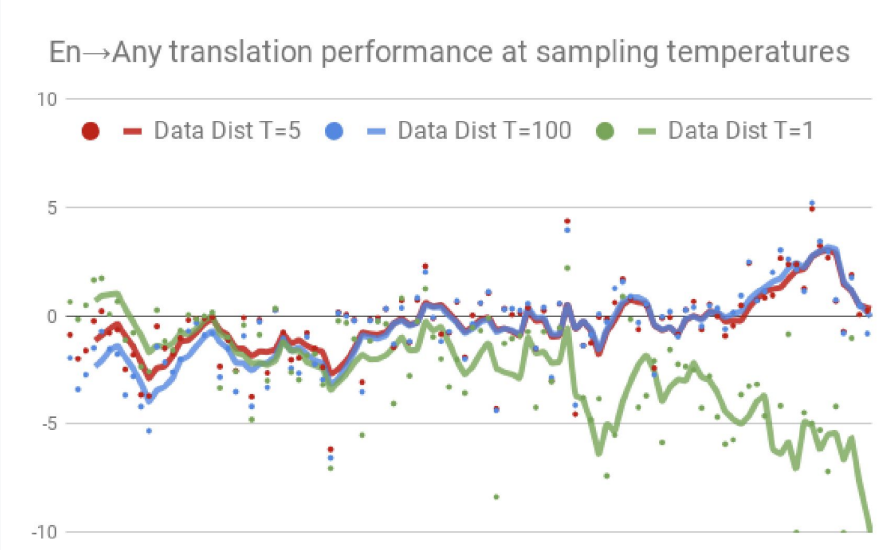
Losses in our best models, but gains in worst models

Pushing the limits of model size

Open Question: Can we train one model for all languages?

Lessons Learned:

- Oversampling works & doesn't work
- Solving the Multi-Task problem helps



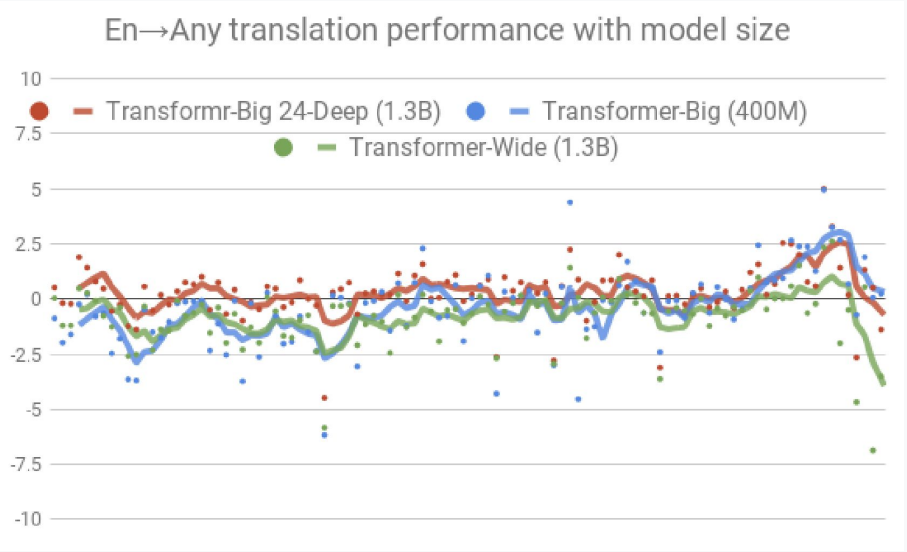
We can recover our losses with smart data ordering

Pushing the limits of model size

Open Question: Can we train one model for all languages?

Lessons Learned:

- Oversampling works & doesn't work
- Solving the Multi-Task problem helps
- Really Deep models help (with a cost)



More layers works best. Requires Transparent Attention and more decoding time

Pushing the limits of model size

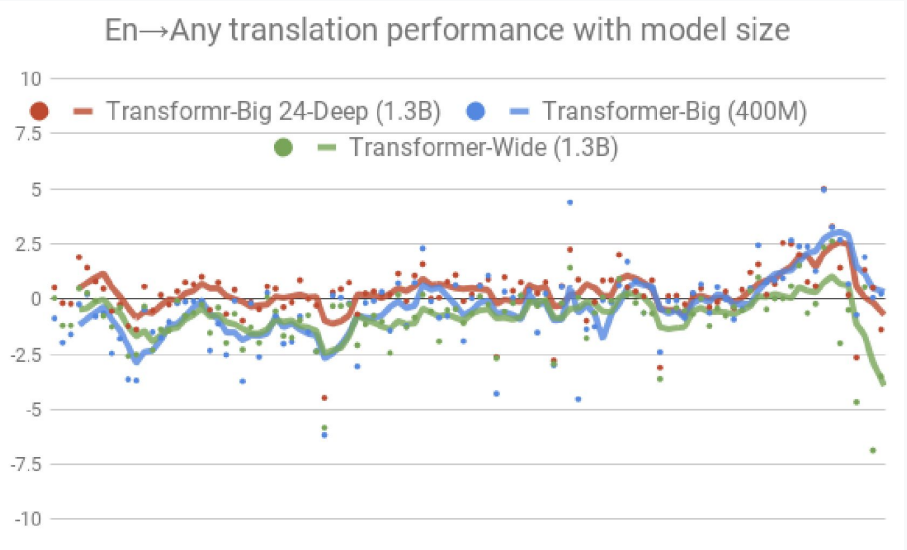
Open Question: Can we train one model for all languages?

Lessons Learned:

- Oversampling works & doesn't work
- Solving the Multi-Task problem helps
- Really Deep models help (with a cost)

Conclusions:

- Need more multi-task approaches
- Can we learn from monolingual data?
- Larger models push theoretical and practical limits
- Handling vocabularies poses more and more problems



More layers works best. Requires Transparent Attention and more decoding time

Translating on the fly

Open Question: Can we make a simultaneous machine translation system?

Paper: Monotonic Infinite Lookback Attention for Simultaneous Machine Translation

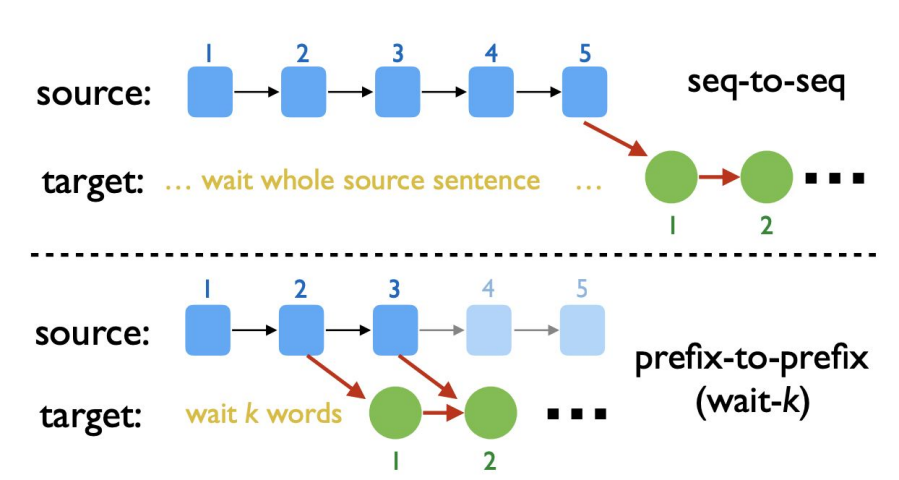
Baidu's STACL Model

Translating on the fly

Open Question: Can we make a simultaneous machine translation system?

Problems:

- No adaptation to context and translation



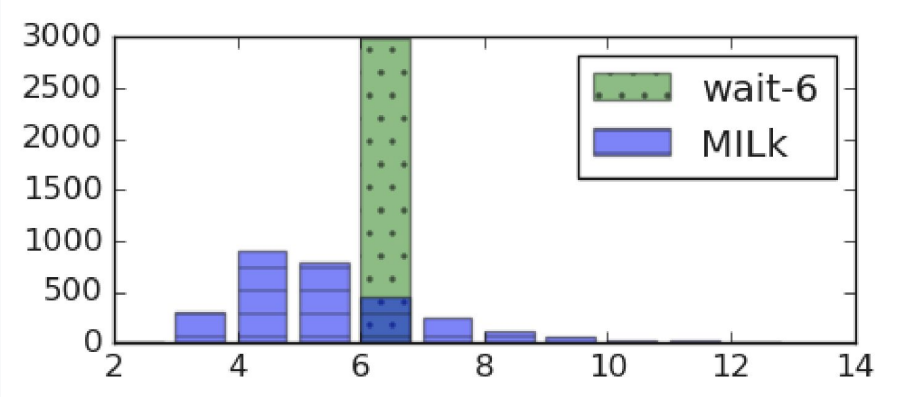
The general prefix to prefix model, with fixed delays

Translating on the fly

Open Question: Can we make a simultaneous machine translation system?

Problems:

- No adaptation to context and translation



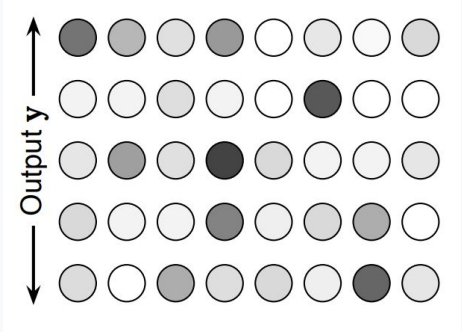
Fixed Delays don't model the data well

Translating on the fly

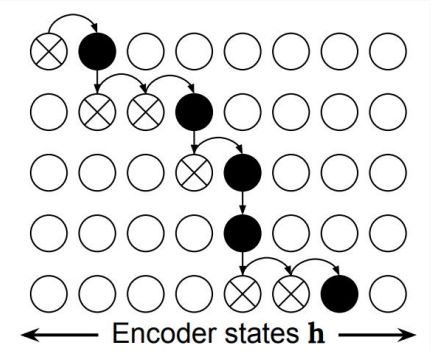
Open Question: Can we make a simultaneous machine translation system?

Problems:

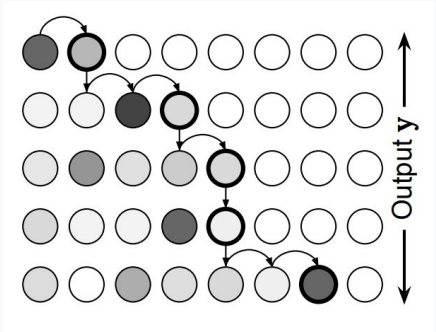
- No adaptation to context and translation
- Attention is Hard



Not Possible



Ignores Information



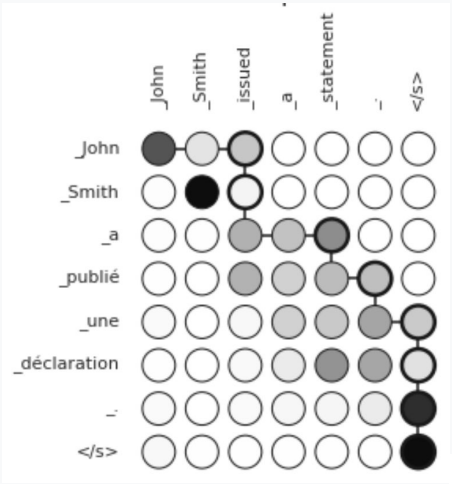
Could learn to wait forever

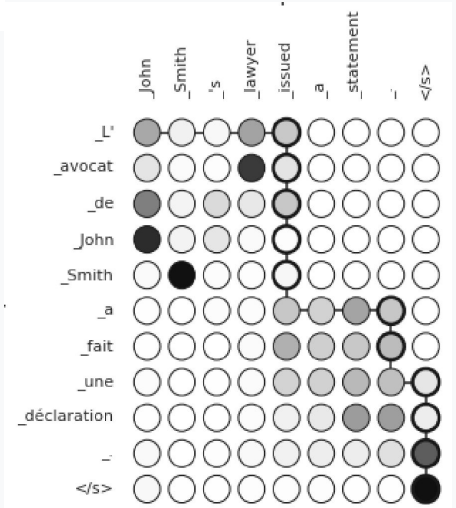
Translating on the fly

Open Question: Can we make a simultaneous machine translation system?

Lessons Learned:

- Blend fixed and soft attention



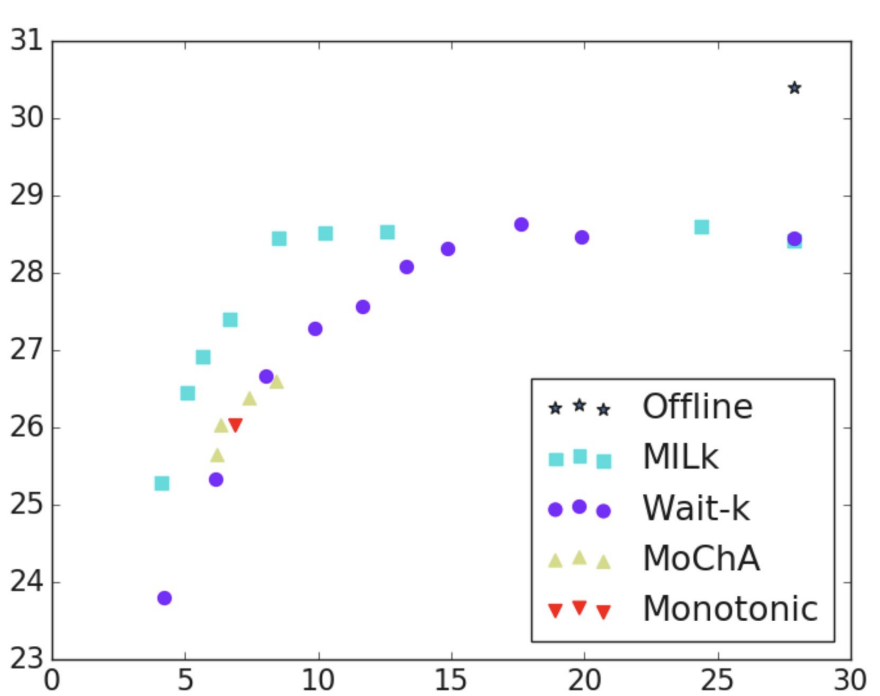


Translating on the fly

Open Question: Can we make a simultaneous machine translation system?

Lessons Learned:

- Blend fixed and soft attention
- Including delay in the loss teaches how to wait



Can get great quality without too much delay

Translating on the fly

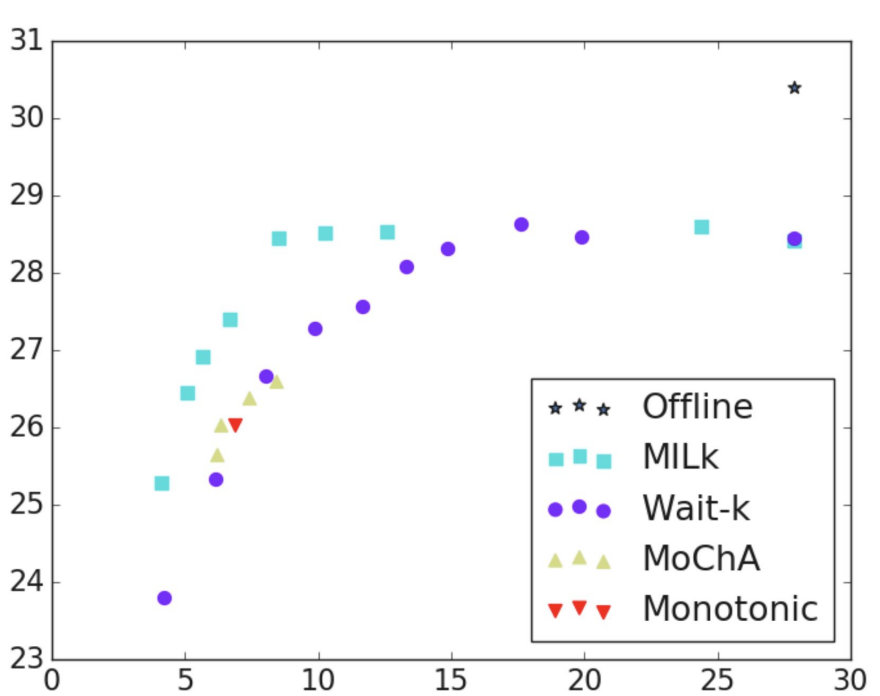
Open Question: Can we make a simultaneous machine translation system?

Lessons Learned:

- Blend fixed and soft attention
- Including delay in the loss teaches how to wait

Conclusions:

- More work to be done
- Small tweaks can lead to big improvements



Can get great quality without too much delay

Working with non-ideal data

Open Question: Our data has noise. Our data doesn't cover all domains. Can we fix that?

Papers:

Denoising Neural Machine Translation Training with Trusted Data and Online Data Selection

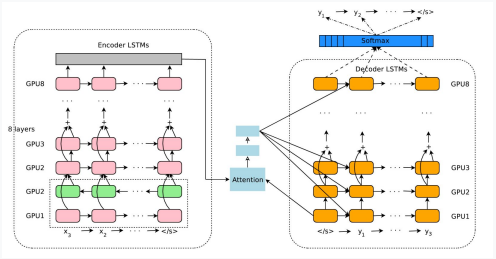
Dynamically Composing Domain-Data Selection with Clean-Data Selection by "Co-Curricular Learning" for Neural Machine Translation

Working with non-ideal data

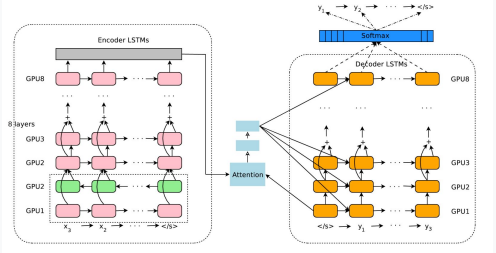
Open Question: Our data has noise. Our data doesn't cover all domains. Can we fix that?

Lessons Learned:

- Cleaning our data improves quality



A small base model on noisy data



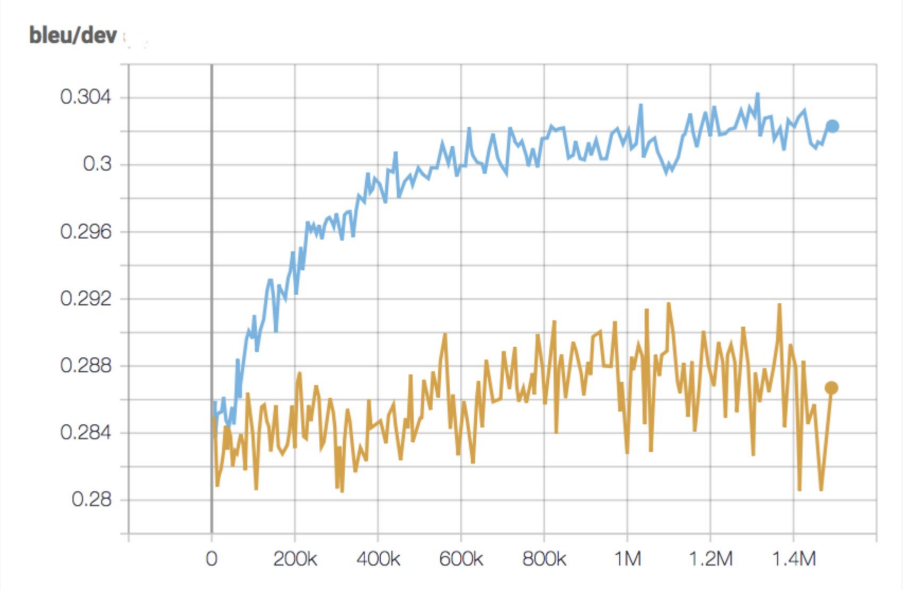
A small clean adapted model on cleaner data

Working with non-ideal data

Open Question: Our data has noise. Our data doesn't cover all domains. Can we fix that?

Lessons Learned:

- Cleaning our data improves quality



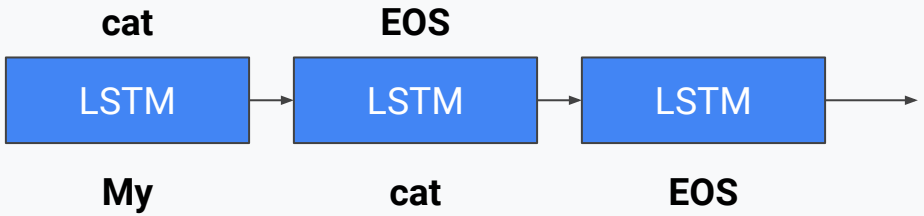
With the right models, we can clean up some noise

Working with non-ideal data

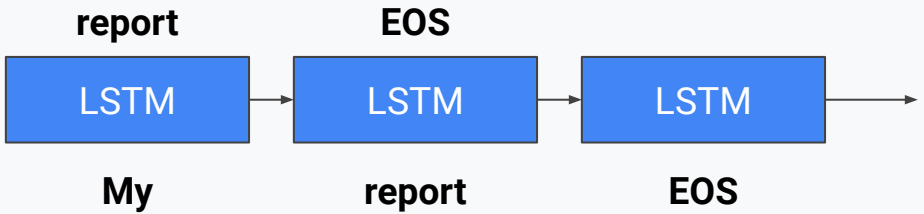
Open Question: Our data has noise. Our data doesn't cover all domains. Can we fix that?

Lessons Learned:

- Cleaning our data improves quality
- Use same approach for domain selection



Base Language Model on general data



Adapted Language Model on domain data

Working with non-ideal data

Open Question: Our data has noise. Our data doesn't cover all domains. Can we fix that?

Lessons Learned:

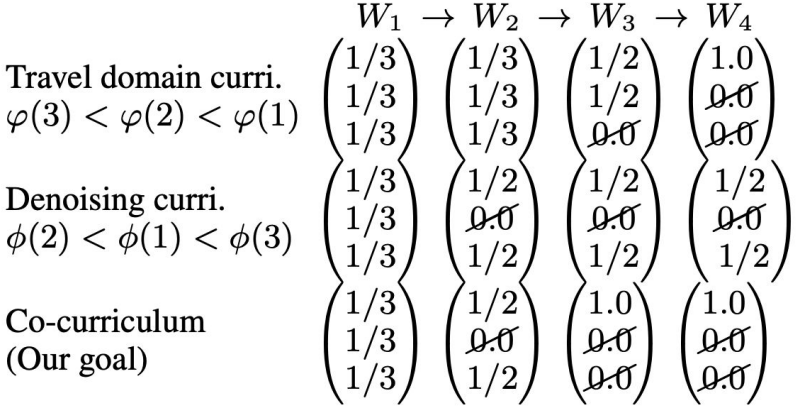
- Cleaning our data improves quality
- Use same approach for domain selection
- A good schedule blends the two approaches

3 en→zh sentence pairs:

1 (en) Where is the train station?
(zh-gloss) TRAIN STATION IS WHERE?

2 (en) Id like to have two window seats.
(zh-gloss) PLEASE BOOK ME TWO WINDOW SEATS.

3 (en) It usually infects people older than 60.
(zh-gloss) PEOPLE OLDER THAN 60 USUALLY ARE INFECTED BY IT.



Preferring different data, at different training steps

Working with non-ideal data

Open Question: Our data has noise. Our data doesn't cover all domains. Can we fix that?

Lessons Learned:

- Cleaning our data improves quality
- Use same approach for domain selection
- A good schedule blends the two approaches

Conclusions:

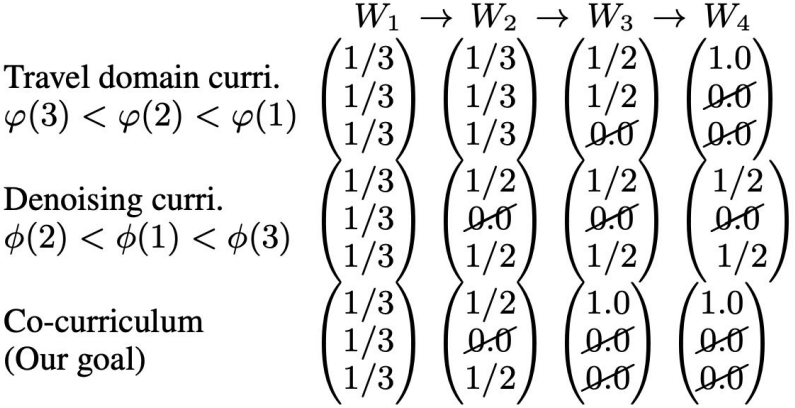
- Possible to learn from multiple signals at once
- Simplifies data needed

3 en→zh sentence pairs:

1
(en) Where is the train station?
(zh-gloss) TRAIN STATION IS WHERE?

2
(en) Id like to have two window seats.
(zh-gloss) PLEASE BOOK ME TWO WINDOW SEATS.

3
(en) It usually infects people older than 60.
(zh-gloss) PEOPLE OLDER THAN 60 USUALLY ARE INFECTED BY IT.



Preferring different data, at different training steps

Using Embeddings to mine data

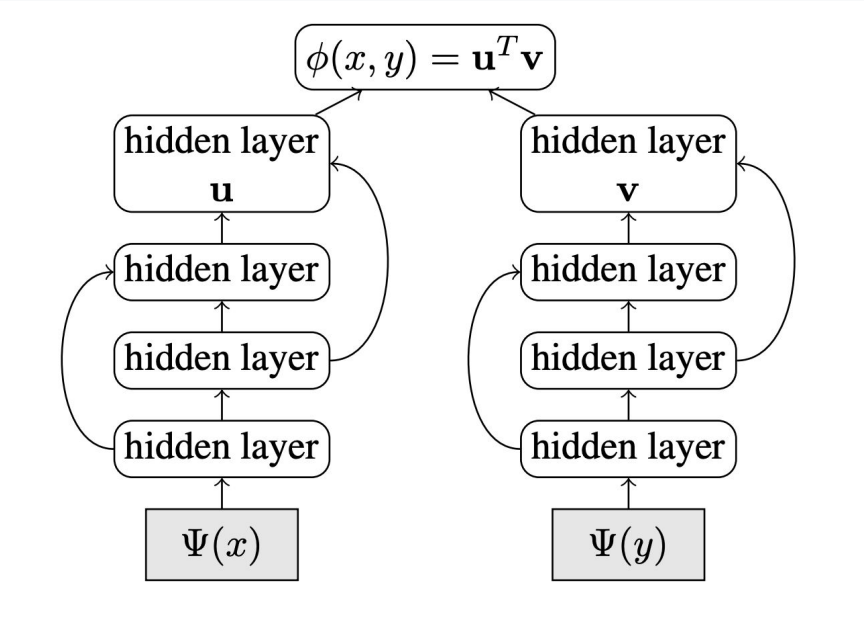
Open Question: Can semantic embeddings improve discovery of parallel data?

Paper: Effective Parallel Corpus Mining using Bilingual Sentence Embeddings

Using Embeddings to mine data

Open Question: Can semantic embeddings improve discovery of parallel data?

Paper: Effective Parallel Corpus Mining using Bilingual Sentence Embeddings



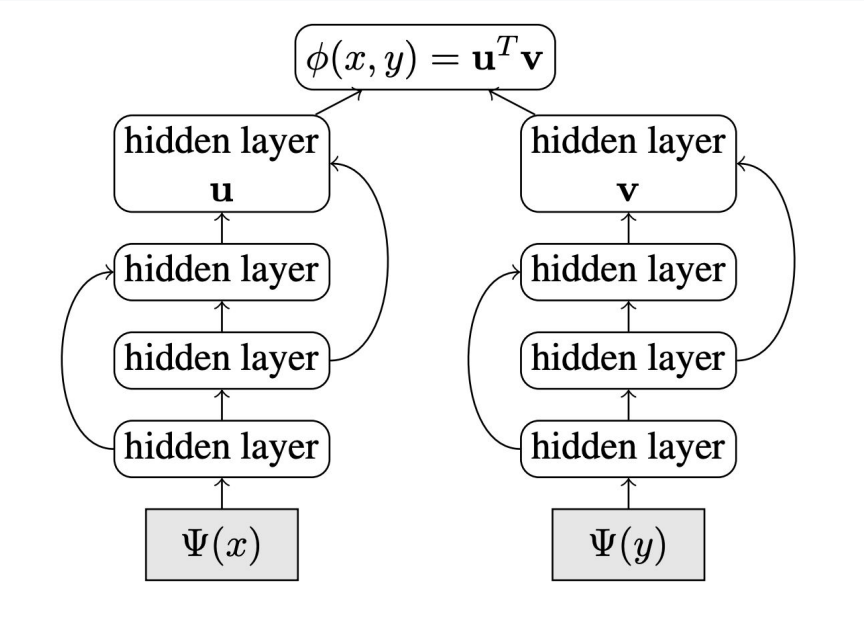
Dual Encoder for Semantic Representations

Using Embeddings to mine data

Open Question: Can semantic embeddings improve discovery of parallel data?

Paper: Effective Parallel Corpus Mining using Bilingual Sentence Embeddings

Limitations: Translation data need to be more than similar



Dual Encoder for Semantic Representations

Using Embeddings to mine data

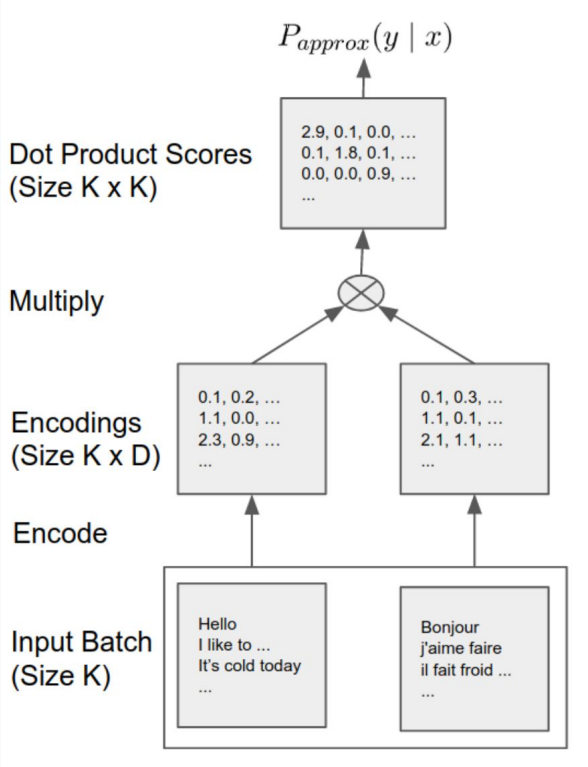
Open Question: Can semantic embeddings improve discovery of parallel data?

Paper: Effective Parallel Corpus Mining using Bilingual Sentence Embeddings

Limitations: Translation data need to be more than similar

Possible Fix:

Learn from negative samples



Incorporating Real and Sampled negatives

Using Embeddings to mine data

Open Question: Can semantic embeddings improve discovery of parallel data?

Paper: Effective Parallel Corpus Mining using Bilingual Sentence Embeddings

Limitations: Translation data need to be more than similar

Possible Fix:

Learn from negative samples

Negative Selection Approach	en-fr		
	P@1	P@3	P@10
Random Negative	34.83	47.99	61.20
Random Negative (Augmented)	36.51	49.07	61.37
(20) Hard Negative	48.90	62.26	73.03

A small number of real negatives makes major gains

Using Embeddings to mine data

Open Question: Can semantic embeddings improve discovery of parallel data?

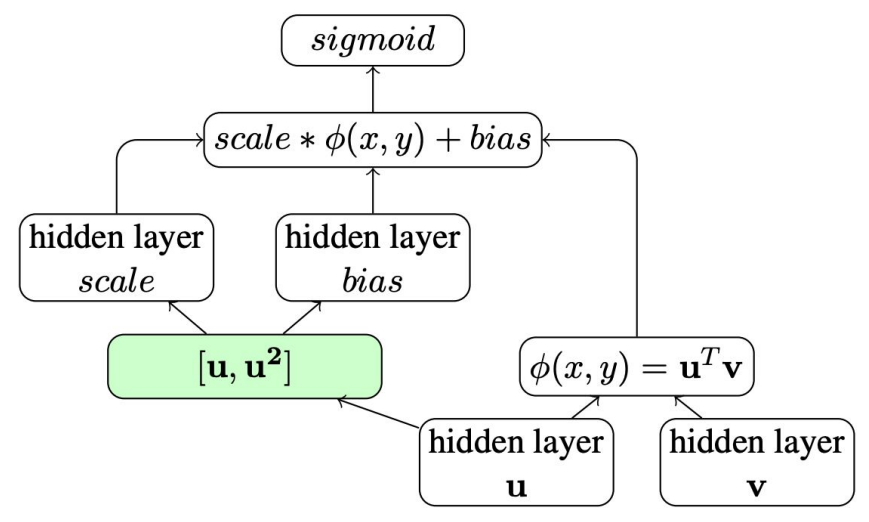
Paper: Effective Parallel Corpus Mining using Bilingual Sentence Embeddings

Limitations: Translation data need to be more than similar

Possible Fix:

Learn from negative samples

Rank documents with sentence matches



Rescaling for comparable ranking of neighbors

Using Embeddings to mine data

Open Question: Can semantic embeddings improve discovery of parallel data?

Paper: Effective Parallel Corpus Mining using Bilingual Sentence Embeddings

Limitations: Translation data need to be more than similar

Possible Fix:

Learn from negative samples

Rank documents with sentence matches

	en-fr (wmt14)	en-es (wmt13)
Mined sentence-level	29.63	29.03
Mined document-level	30.05	27.09
Oracle	30.96	28.81

Document level searching helps

Using Embeddings to mine data

Open Question: Can semantic embeddings improve discovery of parallel data?

Paper: Effective Parallel Corpus Mining using Bilingual Sentence Embeddings

Limitations: Translation data need to be more than similar

Possible Fix:

Learn from negative samples

Rank documents with sentence matches

Matching method	en-fr	en-es
Alignment Counts	82.1	85.1
Our approach Eq. (3)	89.0	90.4
Uszkoreit et al. (2010)	93.4	94.4

Gains remain with feature engineered approach

Using Embeddings to mine data

Open Question: Can semantic embeddings improve discovery of parallel data?

Paper: Effective Parallel Corpus Mining using Bilingual Sentence Embeddings

Limitations: Translation data need to be more than similar

Possible Fix:

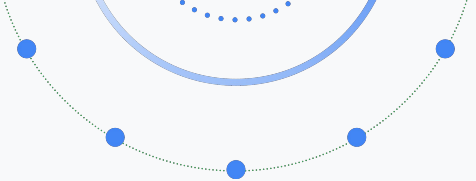
Learn from negative samples

Rank documents with sentence matches

Results: Negatives help. More to be done on document level retrieval.

Matching method	en-fr	en-es
Alignment Counts	82.1	85.1
Our approach Eq. (3)	89.0	90.4
Uszkoreit et al. (2010)	93.4	94.4

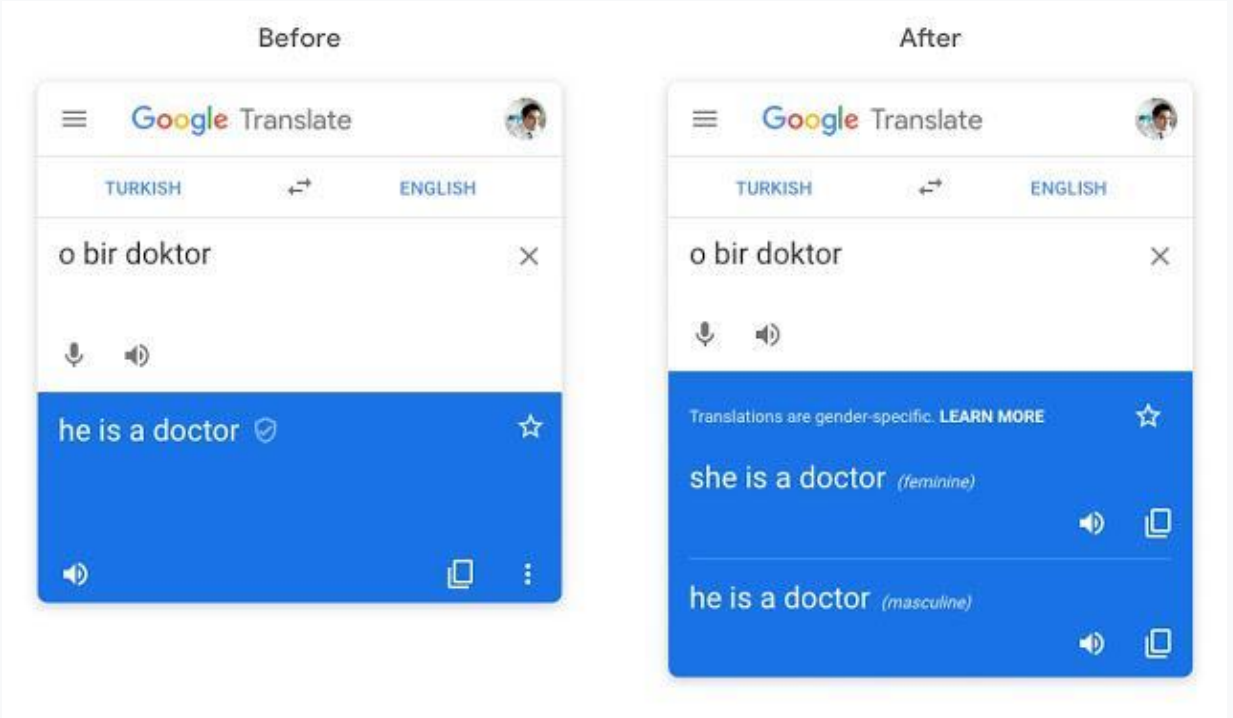
Gains remain with feature engineered approach



Innovations Lead to Real Impact

Research to Users

Gender in Translation:
Making models fair across languages



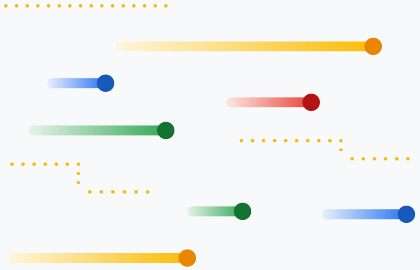
Providing Gender-Specific Translations in Google Translate

Research to Users

Gender in Translation:
Making models fair across languages

Keeping your voice:
Speech to Speech with your tone and style

Introducing Translatotron: An End-to-End Speech-to-Speech Translation Model



Final Questions?

Thank you