Beyond Parallel Corpora

Philipp Koehn

26 October 2023

data and machine learning

Supervised and Unsupervised

- We framed machine translation as a supervised machine learning task
  - training examples with labels
  - here: input sentences with translation
  - structured prediction: output has to be constructed in several steps

Supervised and Unsupervised

- We framed machine translation as a supervised machine learning task
  - training examples with labels
  - here: input sentences with translation
  - structured prediction: output has to be constructed in several steps

- Unsupervised learning
  - training examples without labels
  - here: just sentences in the input language
  - we will also look at using just sentences output language

Supervised and Unsupervised

- We framed machine translation as a supervised machine learning task
  - training examples with labels
  - here: input sentences with translation
  - structured prediction: output has to be constructed in several steps

- Unsupervised learning
  - training examples without labels
  - here: just sentences in the input language
  - we will also look at using just sentences output language

- Semi-supervised learning
  - some labeled training data
  - some unlabeled training data (usually more)

Supervised and Unsupervised

- We framed machine translation as a supervised machine learning task
  - training examples with labels
  - here: input sentences with translation
  - structured prediction: output has to be constructed in several steps

- Unsupervised learning
  - training examples without labels
  - here: just sentences in the input language
  - we will also look at using just sentences output language

- Semi-supervised learning
  - some labeled training data
  - some unlabeled training data (usually more)

- Self-training
  - make predictions on unlabeled training data
  - use predicted labeled as supervised translation data

Transfer Learning

- Learning from data similar to our task

Transfer Learning

- Learning from data similar to our task

- Other language pairs
  - first, train a model on different language pair
  - then, train on the targeted language pair
  - or: train jointly on both

Transfer Learning

- Learning from data similar to our task

- Other language pairs
  - first, train a model on different language pair
  - then, train on the targeted language pair
  - or: train jointly on both

- Multi-Task training
  - train on a related task first
  - e.g., part-of-speeh tagging

- Share some or all of the components

using monolingual data

Using Monolingual Data

- Language model
  - trained on large amounts of target language data
  - better fluency of output
- Key to success of statistical machine translation
- Neural machine translation
  - integrate neural language model into model
  - create artificial data with backtranslation

Adding a Language Model

- Train a separate language model

- Add as conditioning context to the decoder

Adding a Language Model

- Train a separate language model
- Add as conditioning context to the decoder
- Recall state progression in the decoder
  - decoder state s_i
  - embedding of previous output word Ey_{i-1}
  - input context c_i

$$s_i = f(s_{i-1},\ Ey_{i-1}, c_i)$$

Adding a Language Model

- Train a separate language model

- Add as conditioning context to the decoder

- Recall state progression in the decoder
  - decoder state s_i
  - embedding of previous output word Ey_{i-1}
  - input context c_i

$$s_i = f(s_{i-1},\ Ey_{i-1}, c_i)$$

- Add hidden state of neural language model s_i^{LM}

$$s_i = f(s_{i-1},\ Ey_{i-1}, c_i, s_i^{\mathsf{LM}})$$

Adding a Language Model

- Train a separate language model

- Add as conditioning context to the decoder

- Recall state progression in the decoder
  - decoder state s_i
  - embedding of previous output word Ey_{i-1}
  - input context c_i

$$s_i = f(s_{i-1},\ Ey_{i-1}, c_i)$$

- Add hidden state of neural language model s_i^{LM}

$$s_i = f(s_{i-1},\ Ey_{i-1}, c_i, s_i^{\text{LM}})$$

- Pre-train language model

- Leave its parameters fixed during translation model training

Refinements

- Balance impact of language model vs. translation model

Refinements

- Balance impact of language model vs. translation model
- Learn a scaling factor (gate)

$$\text{gate}_i^{\mathsf{LM}} = f(s_i^{\mathsf{LM}})$$

Refinements

- Balance impact of language model vs. translation model
- Learn a scaling factor (gate)

$$\text{gate}_i^{\mathsf{LM}} = f(s_i^{\mathsf{LM}})$$

- Use it to scale values of language model state

$$\bar{s}_i^{\mathsf{LM}} = \text{gate}_i^{\mathsf{LM}} \times s_i^{\mathsf{LM}}$$

Refinements

- Balance impact of language model vs. translation model
- Learn a scaling factor (gate) $\text{gate}_i^{\textsf{LM}} = f(s_i^{\textsf{LM}})$
- Use it to scale values of language model state

$$\bar{s}_i^{\textsf{LM}} = \text{gate}_i^{\textsf{LM}} \times s_i^{\textsf{LM}}$$

- Use this scaled language model state for decoder state

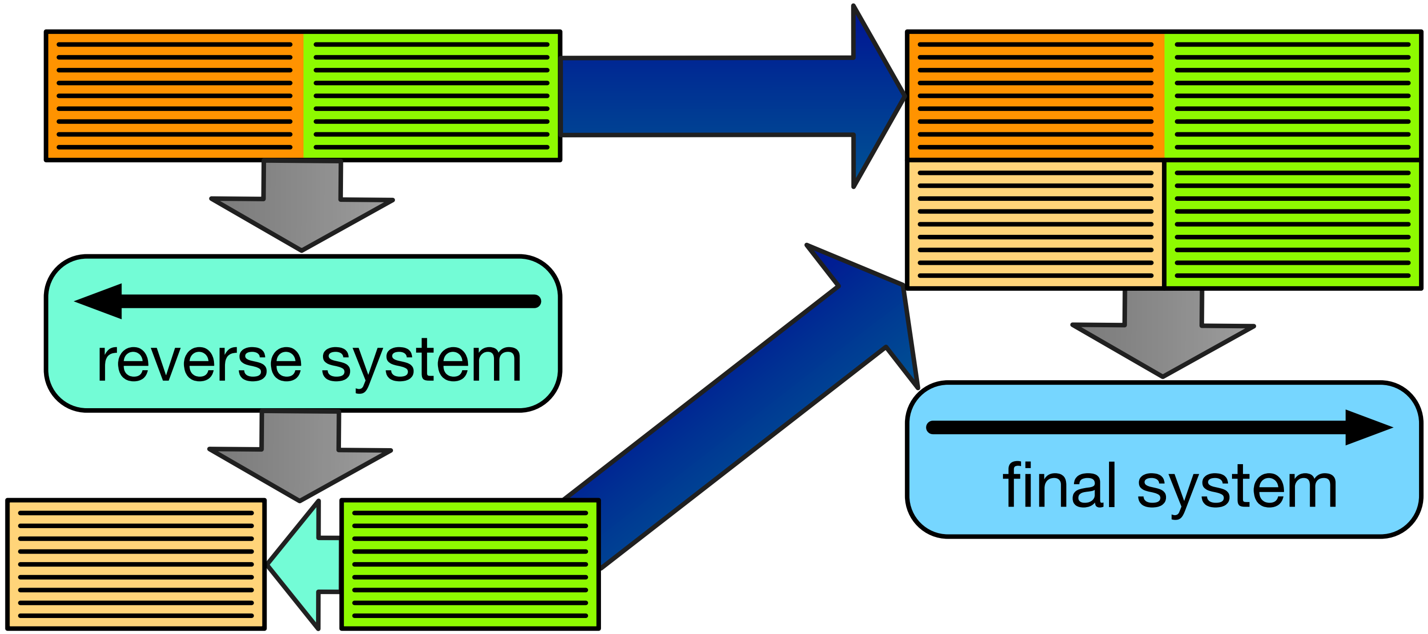
$$s_i = f(s_{i-1},\ Ey_{i-1}, c_i, \bar{s}_i^{\textsf{LM}})$$

Back Translation

- Monolingual data is parallel data that misses its other half

Back Translation

- Monolingual data is parallel data that misses its other half
- Let's synthesize that half



Back Translation

- Steps
  1. train a system in reverse language translation
  2. use this system to translate target side monolingual data
     $\rightarrow$ synthetic parallel corpus
  3. combine generated synthetic parallel data with real parallel data to build the final system
- Roughly equal amounts of synthetic and real data
- Useful method for domain adaptation

Iterative Back Translation

- Quality of backtranslation system matters

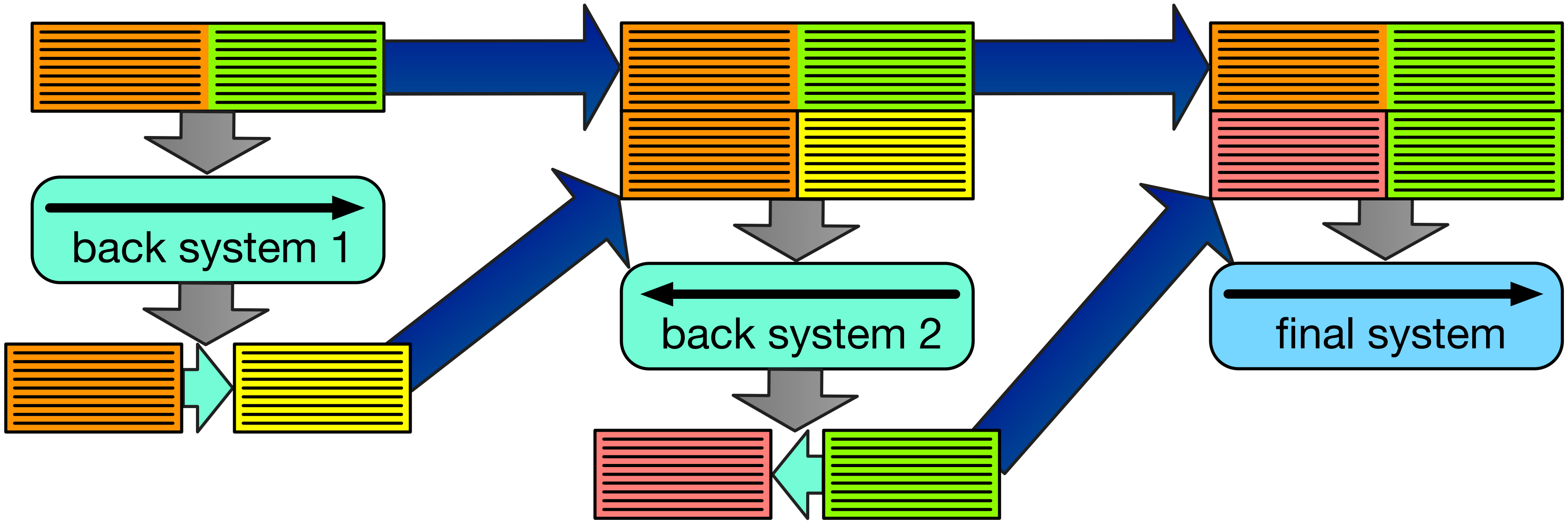
Iterative Back Translation

- Quality of backtranslation system matters

- Build a better backtranslation system ... with backtranslation

Iterative Back Translation

- Quality of backtranslation system matters
- Build a better backtranslation system ... with backtranslation



Iterative Back Translation

- Example

German–English	Back	Final
no back-translation	-	29.6
*10k iterations	10.6	29.6 (+0.0)
*100k iterations	21.0	31.1 (+1.5)
convergence	23.7	32.5 (+2.9)
re-back-translation	27.9	33.6 (+4.0)

* = limited training of back-translation system

Variants

- Copy Target
  - if no good neural machine translation system to start with
  - just copy target language text to the source

- Forward Translation
  - synthesize training data in same direction as training
  - self-training (inferior but sometimes successful)

Round Trip Training

- We could iterate through steps of
  - train system
  - create synthetic corpus

Round Trip Training

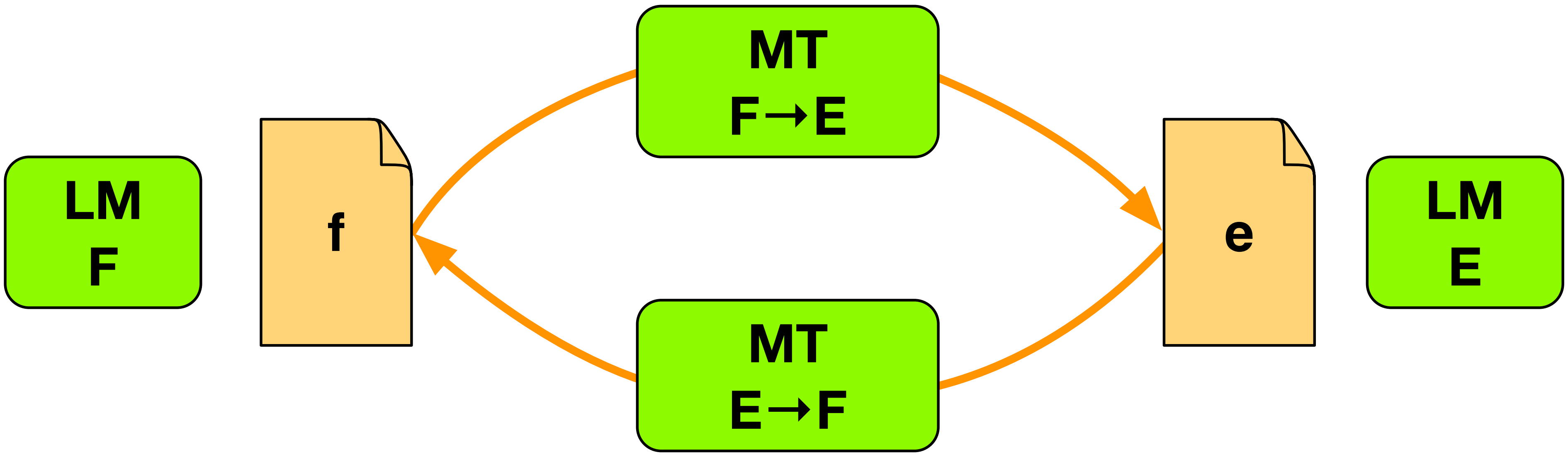
- We could iterate through steps of
  - train system
  - create synthetic corpus
- Dual learning: train models in both directions together
  - translation models $F \rightarrow E$ and $E \rightarrow F$
  - take sentence **f**
  - translate into sentence **e'**
  - translate that back into sentence **f'**
  - training objective: **f** should match **f'**

Round Trip Training

- We could iterate through steps of
  - train system
  - create synthetic corpus
- Dual learning: train models in both directions together
  - translation models $F \rightarrow E$ and $E \rightarrow F$
  - take sentence **f**
  - translate into sentence **e'**
  - translate that back into sentence **f'**
  - training objective: **f** should match **f'**
- Setup could be fooled by just copying (**e'** = **f**)

  ⇒ score **e'** with a language for language E
  add language model score as cost to training objective

Round Trip Training



Monolingual Pre-Training

- Initial training of neural machine translation model on monolingual data
- Replace some input word sequences with <pad> (30% of words)
- Train model MASKED → TEXT on both source and target text
- Reorder sentences (each training example has 3 sentences)

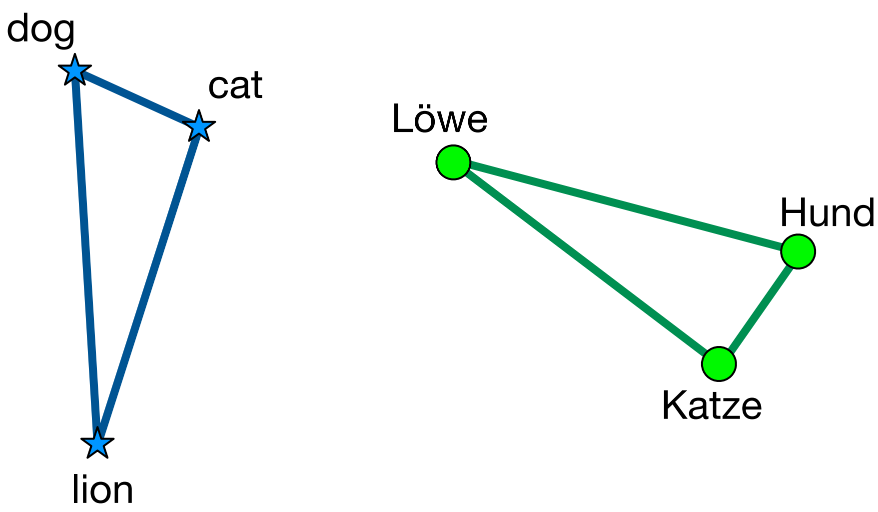
<en> Advanced NLP techniques master class " how <pad> " </s>
3rd <pad> : 18 </s>
Results <pad> 40 of 729

⇓

3rd grade : 18 </s>
Advanced NLP techniques master class " how to with clients " </s>
Results 1 – 40 of 729

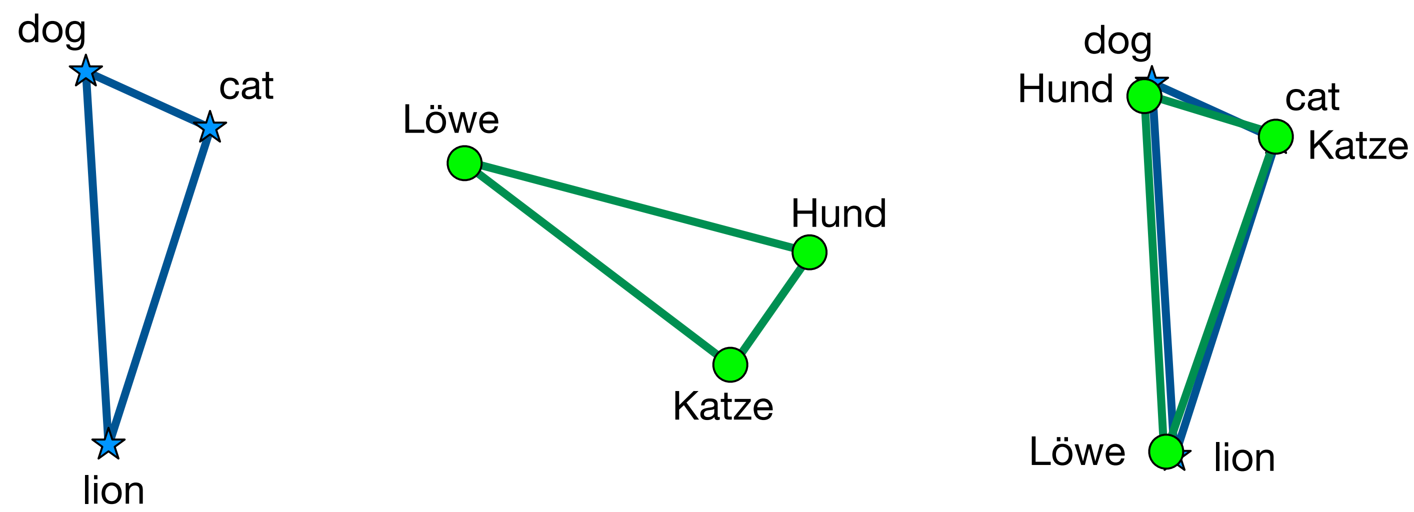
unsupervised machine translation

Monolingual Embedding Spaces



- Embedding spaces for different languages have similar shape
- Intuition: relationship between *dog*, *cat*, and *lion*, independent of language
- How can we rotate the triangle to match up?

Matching Embedding Spaces



- Seed lexicon of identically spelled words, numbers, names
- Adversarial training: discriminator predicts language [Conneau et al., 2018]
- Match matrices with word similarity scores: Vecmap [Artetxe et al., 2018]

Inferred Translation Model

- Translation model
  - induced word translations (nearest neighbors of mapped embeddings)

  $\rightarrow$ statistical phrase translation table (probability $\simeq$ similarity)

- Language model
  - target side monolingual data

  $\rightarrow$ estimate statistical n-gram language model

$\Rightarrow$ Statistical phrase-based machine translation system

Synthetic Training Data

- Create synthetic parallel corpus
  - monolingual text in source language
  - translate with inferred system: translations in target language
- Recall: EM algorithm
  - predict data: generate translation for monolingual corpus
  - predict model: estimate model from synthetic data
  - iterate this process, alternate between language directions
- Increasingly use neural machine translation model to synthesize data

multiple language pairs

Multiple Language Pairs

- There are more than two languages in the world
- We may want to build systems for many language pairs
- Typical: train separate models for each
- Joint training

Multiple Input Languages

- Example
  - German–English
  - French–English
- Concatenate training data
- Joint model benefits from exposure to more English data
- Shown beneficial in low resource conditions
- Do input languages have to be related? (maybe not)

Multiple Output Languages

- Example
  - French–English
  - French–Spanish
- Concatenate training data
- Given a French input sentence, how specify output language?

Multiple Output Languages

- Example
  - French–English
  - French–Spanish

- Concatenate training data

- Given a French input sentence, how specify output language?

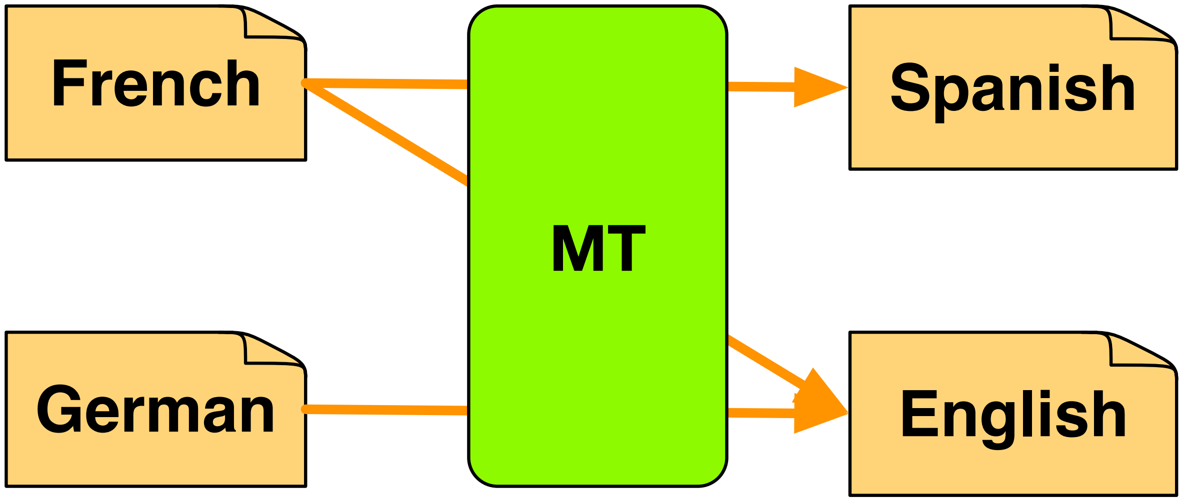
- Indicate output language with special tag

[ENGLISH] *N'y a-t-il pas ici deux poids, deux mesures?*
⇒ *Is this not a case of double standards?*

[SPANISH] *N'y a-t-il pas ici deux poids, deux mesures?*
⇒ *¿No puede verse con toda claridad que estamos utilizando un doble rasero?*

Zero Shot Translation

- Example
  - German–English
  - French–English
  - French–Spanish
- We want to translate
  - German–Spanish



Zero Shot

- Train on
  - German–English
  - French–English
  - French–Spanish
- Specify translation

[SPANISH] *Messen wir hier nicht mit zweierlei Maß?*

$\Rightarrow$ *¿No puede verse con toda claridad que estamos utilizando un doble rasero?*

Zero Shot: Hype

Algorithms

Google's AI just created its own universal 'language'

The technology used in Google Translate can identify hidden material between languages to create what's known as interlingua

---

By **MATT BURGESS**

23 Nov 2016

Zero Shot: Reality

Table 5: Portuguese→Spanish BLEU scores using various models.

	Model	Zero-shot	BLEU
(a)	PBMT bridged	no	28.99
(b)	NMT bridged	no	30.91
(c)	NMT Pt→Es	no	31.50
(d)	Model 1 (Pt→En, En→Es)	yes	21.62
(e)	Model 2 (En↔{Es, Pt})	yes	24.75
(f)	Model 2 + incremental training	no	31.77

- Bridged: pivot translation Portuguese → English → Spanish
- Model 1 and 2: Zero shot training
- Model 2 + incremental training: use of some training data in language pair

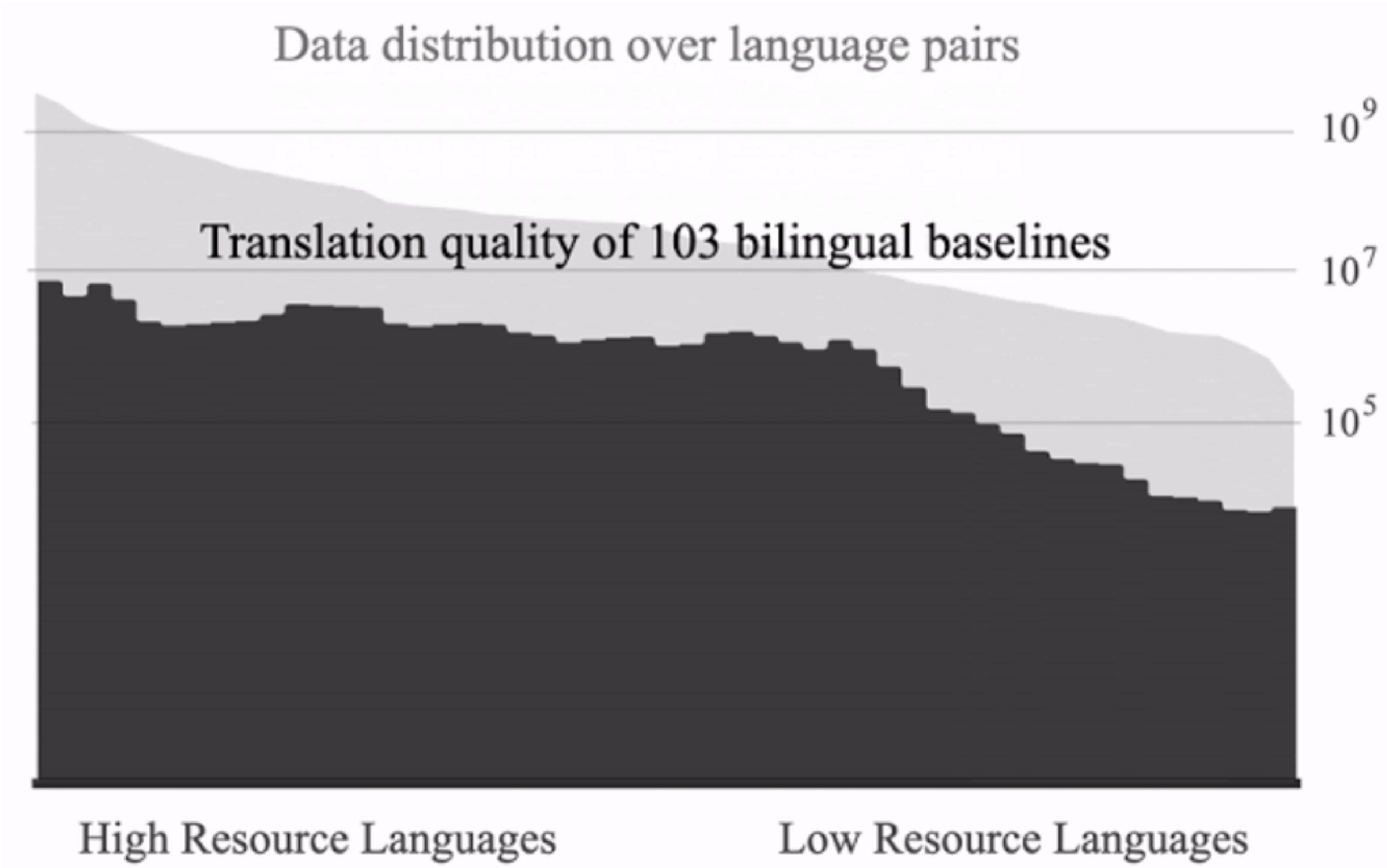
Massively Multilingual Training

- Scaling up multilingual machine translation for more languages
  - many-to-English
  - English-to-many
  - many-to-many

- Mainly motivated by improving low-resource language pairs

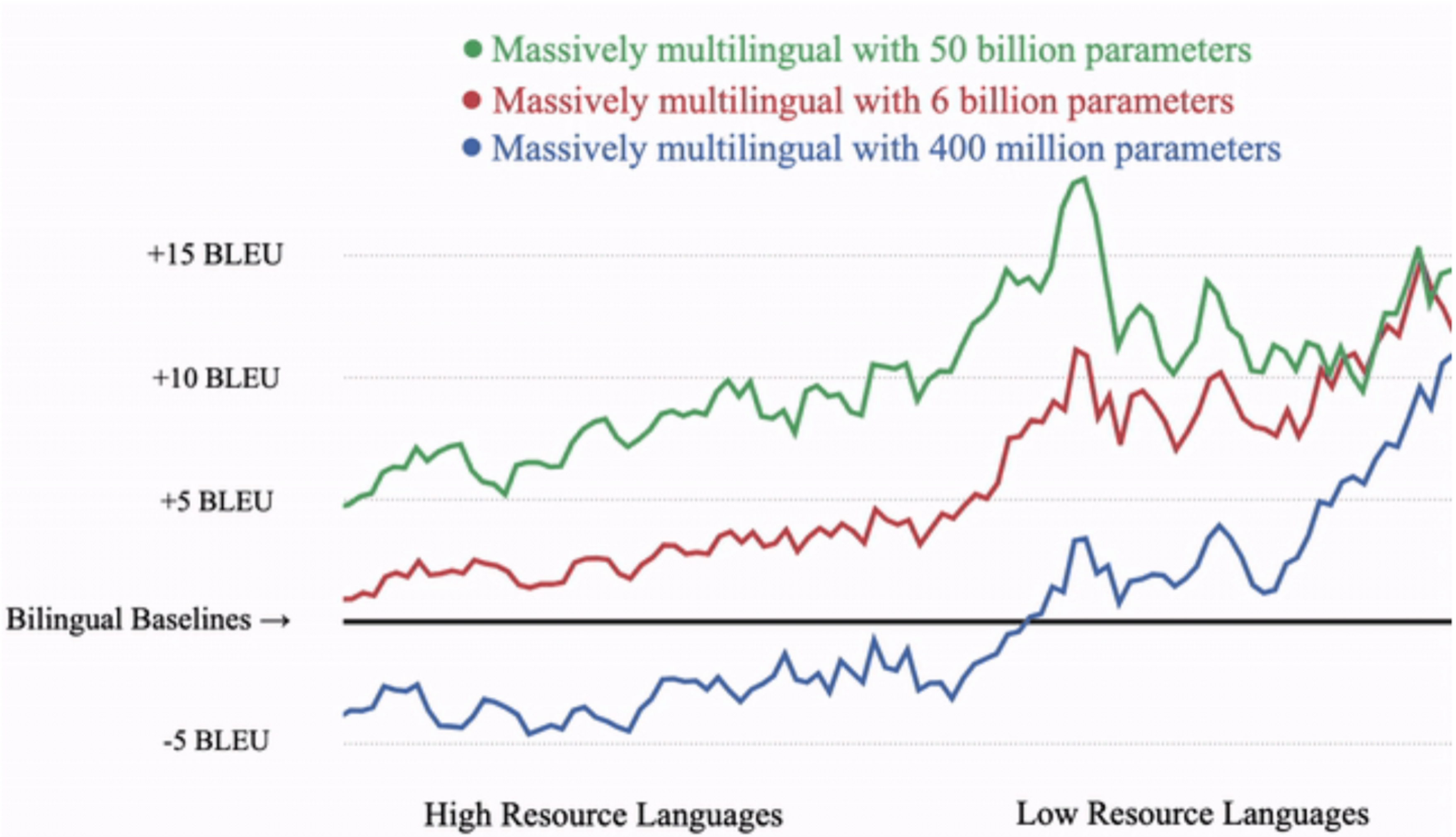
- Move towards larger models

Translation Quality for 103 Languages



(source: Google)

Gains with Multilingual Training



(source: Google)

Romanization

Maximize shared orthography → uroman

Maximize shared vocabulary → BPE

ፊሊፒንሲ	filipenesi	fil ipe nes i
الفلبين	alflbyn	alf lb yn
ফিলিপাইন	Philipaaina	Phil ipa aina
فیلیپین	filipin	fil ipin
Fülöp-szigetek	Fueloep-szigetek	Fue lo ep - szi gete k
Filipina	Filipina	Filipin a
Филиппины	Filippiny	Fili ppi ny
Filibiin	Filibiin	Fili bi in
Filipinas	Filipinas	Filipin as
பிலிப்பைன்ஸ்	pilippains	pil ipp ain s
Philippines	Philippines	Philippines
Philippines	Philippines	Philippines
Filifin	Filifin	Fili fin
Ufilipino	Ufilipino	U fil ipin o
Pilipinas	Pilipinas	Pil ipin as
Filipinler	Filipinler	Filipin ler
Filippin	Filippin	Fili ppi n

(source: USC/ISI)

Many-to-Many

Facebook

Introducing the First AI Model That Translates 100 Languages Without Relying on English

October 19, 2020

By Angela Fan, Research Assistant

- 7.5 billion sentences for 100 languages
  (mined from web-crawled data)
- Model with 15 billion parameters
- Improvements especially for low resource languages

Even Bigger: NLLB (2022)

- No Language Left Behind: 200 languages
- Hand-translated test set: Flores-200
- Uses diverse data sources
  - public parallel data
  - translations created by professional translators
  - sentence pairs based on sentence embedding similarity
  - monolingual data for
    * monolingual pre-training
    * back-translation
    * self-training
- Models of different scale (up to 54B parameters), publicly released

Different Amounts of Data per Language

- High-resource language pairs are undertrained
- Low-resource language pairs are overtrained

Different Amounts of Data per Language

- High-resource language pairs are undertrained
- Low-resource language pairs are overtrained

$\Rightarrow$ Oversampling low resource language pairs
Data selection probability p_l for language pair l based on corpus sizes D_k

$$p_l = (D_l / \sum_k D_k)^{1/T}$$

Different Amounts of Data per Language

- High-resource language pairs are undertrained
- Low-resource language pairs are overtrained

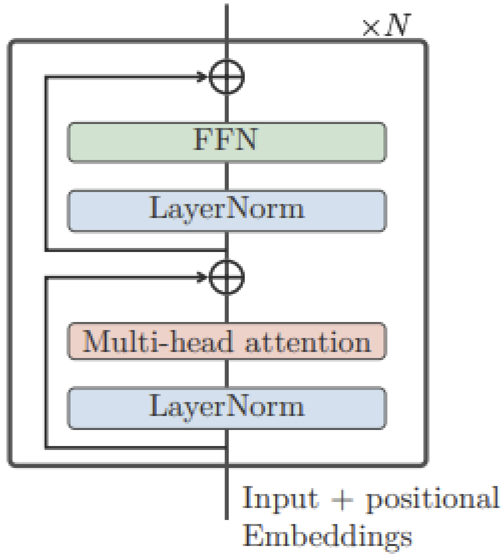
⇒ Oversampling low resource language pairs
Data selection probability p_l for language pair l based on corpus sizes D_k

$$p_l = (D_l / \sum_k D_k)^{1/T}$$

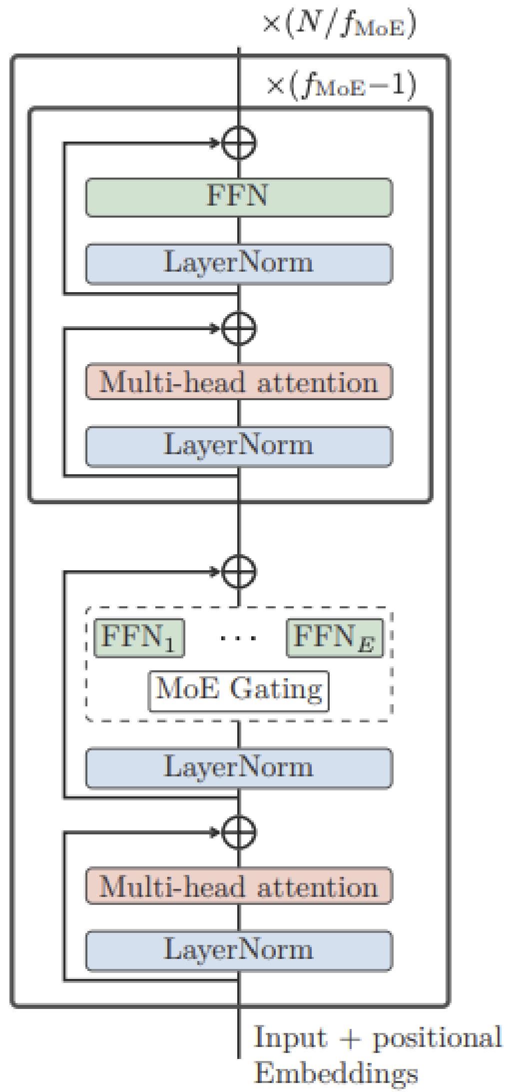
- Curriculum training: adding low-resource data only in later training stages

Mixture of Experts

- Conditional compute
- Gating mechanism decides which FF step to utilize
- Allows scaling to many more parameters without increasing computational cost



(a) Dense Transformer



(b) MoE Transformer