Beyond Parallel Corpora

Philipp Koehn

29 October 2020

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

- 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

- 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

- 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}

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

- Pre-train language model

- Leave its parameters fixed during translation model training

- Balance impact of language model vs. translation model

- Learn a scaling factor (gate)
  $$\text{gate}_i^{\text{LM}} = f(s_i^{\text{LM}})$$

- Use it to scale values of language model state
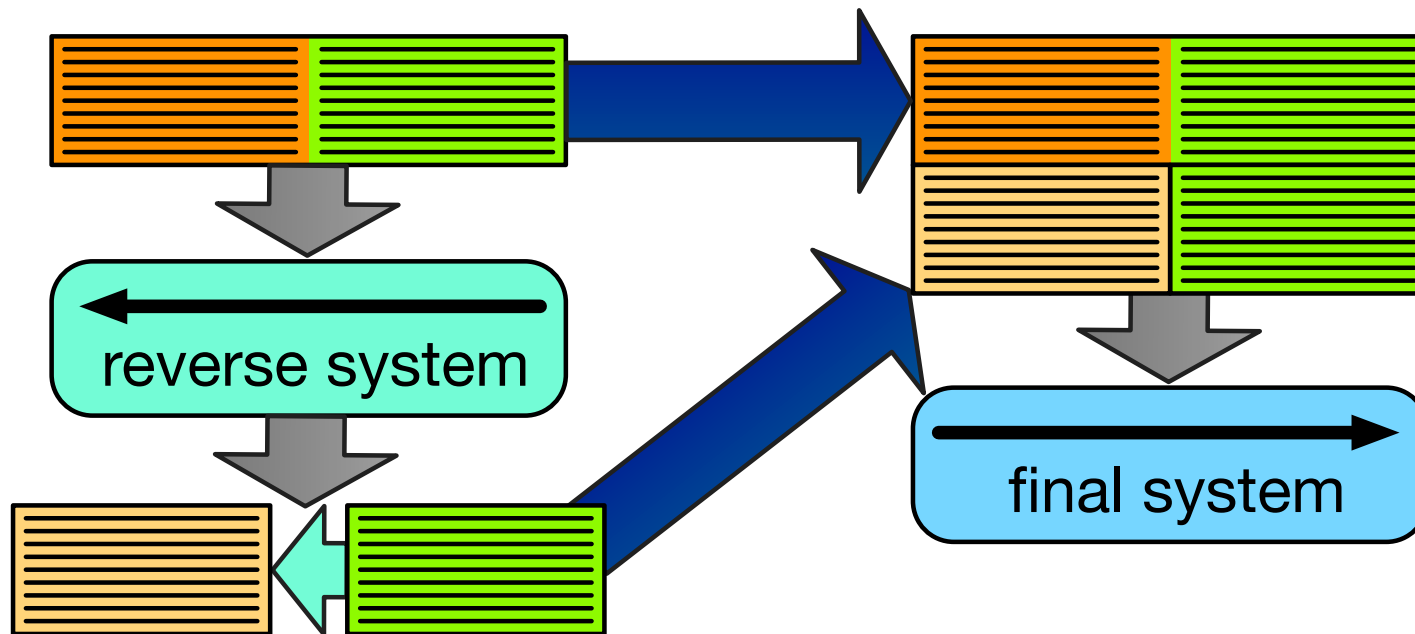 $$\bar{s}_i^{\text{LM}} = \text{gate}_i^{\text{LM}} \times s_i^{\text{LM}}$$

- Use this scaled language model state for decoder state
  $$s_i = f(s_{i-1},\ Ey_{i-1}, c_i, \bar{s}_i^{\text{LM}})$$

- 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▮

- Build a better backtranslation system ... with backtranslation▮
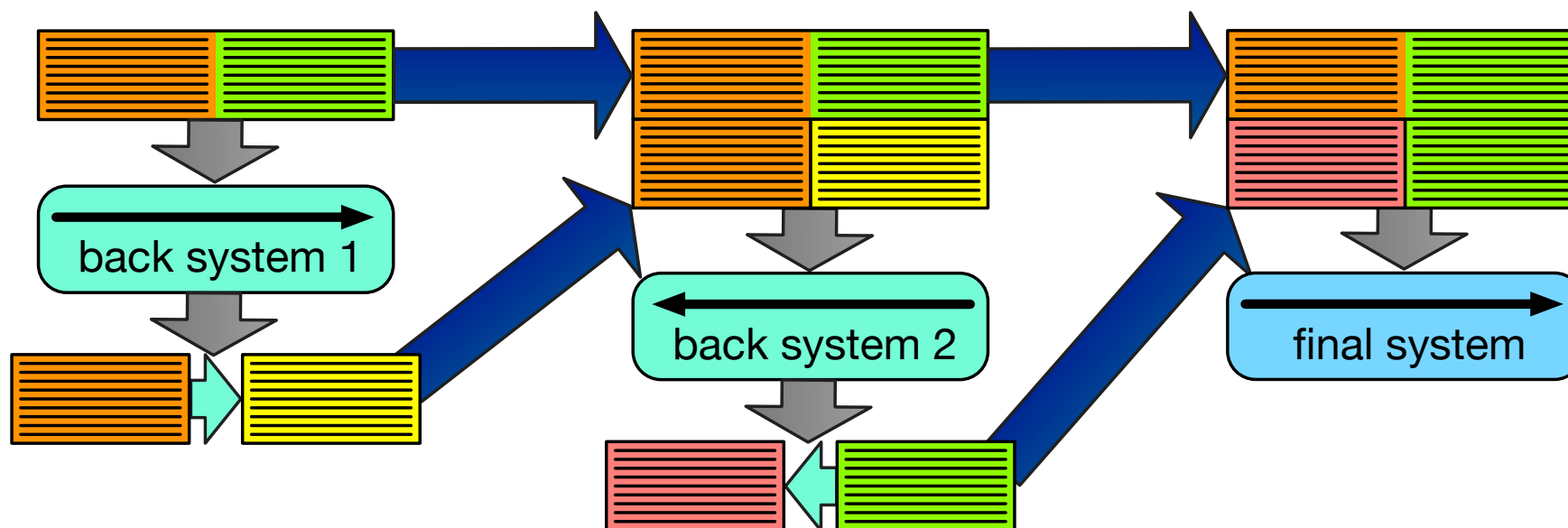
Iterative Back Translation

- 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

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**)

  $\Rightarrow$ score **e'** with a language for language E
  add language model score as cost to training objective
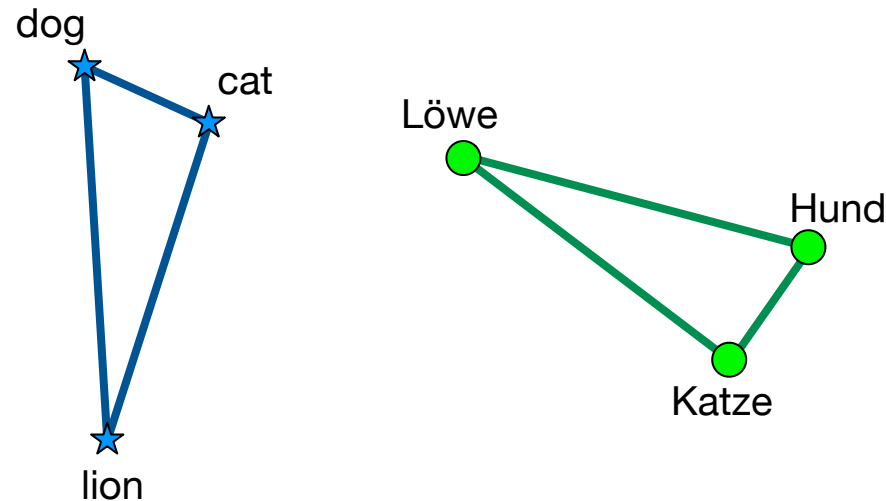
Round Trip Training
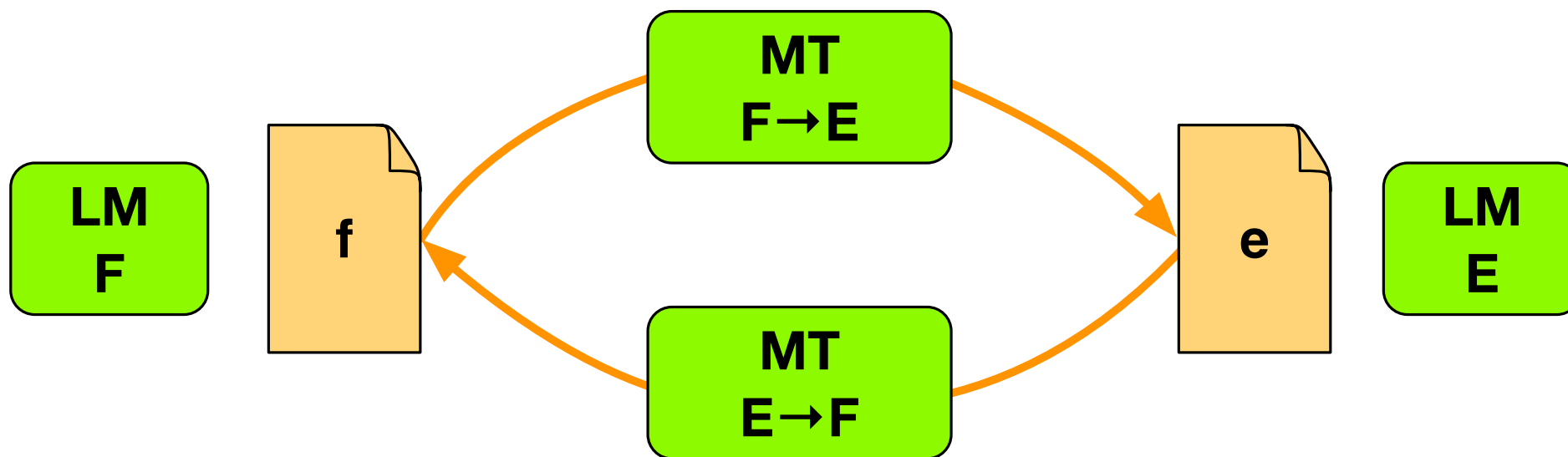
- Copy Target

  - 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)

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
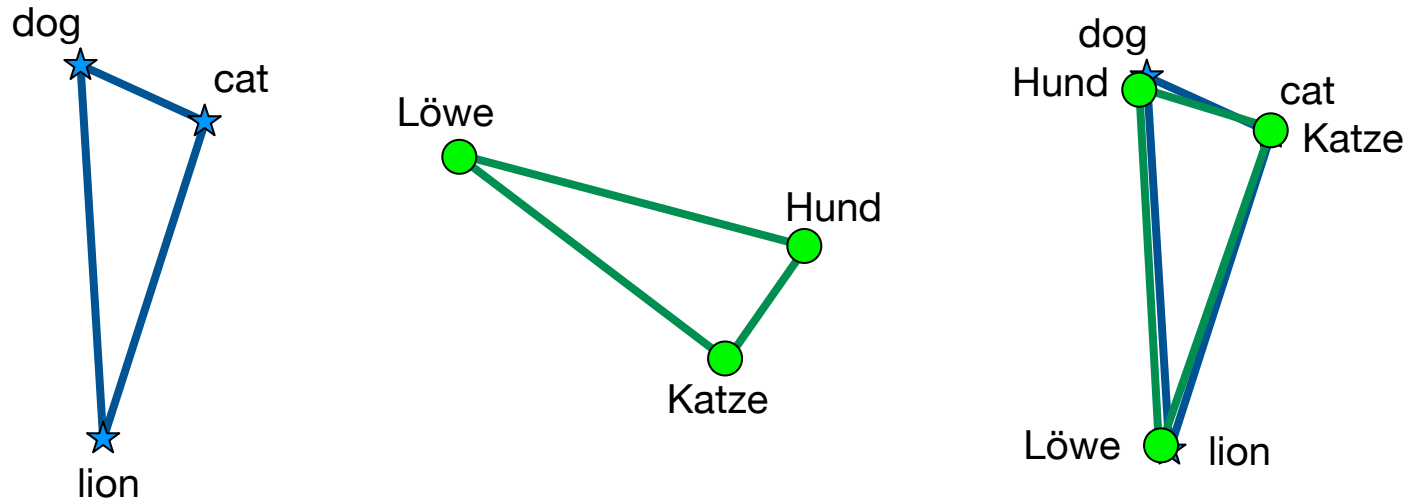
- Seed lexicon of identically spelled words, numbers, names

- Adversarial training: discriminator predicts language [Conneau et al., 2018]

- Match matrices with word similarity scores: Vecmap [Artetxe et al., 2018]

- 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

- 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

- 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)

- 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?*
$\Rightarrow$ *Is this not a case of double standards?*

[SPANISH] *N'y a-t-il pas ici deux poids, deux mesures?*
$\Rightarrow$ *¿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
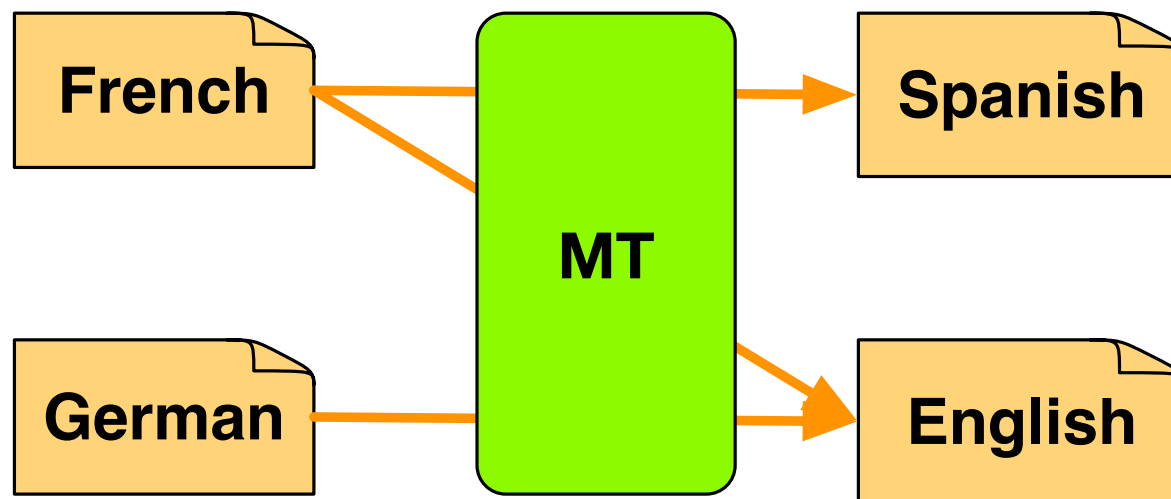
Zero Shot

- Train on

  - German–English
  - French–English
  - French–Spanish

- Specify translation

[SPANISH] *Messen wir hier nicht mit zweierlei Maß?*
    ⇒ ¿No puede verse con toda claridad que estamos utilizando un doble rasero?

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

Sharing Components

- So far: generic neural machine translation model

- Maybe better: separate systems with shared components

  - encoder shared in models with same input language.
  - decoder shared in models with same output language.
  - attention mechanism shared in all models

- Sharing = same parameters, updates from any language pair training

- No need to mark output language

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

Data distribution over language pairs

Translation quality of 103 bilingual baselines

High Resource Languages   Low Resource Languages
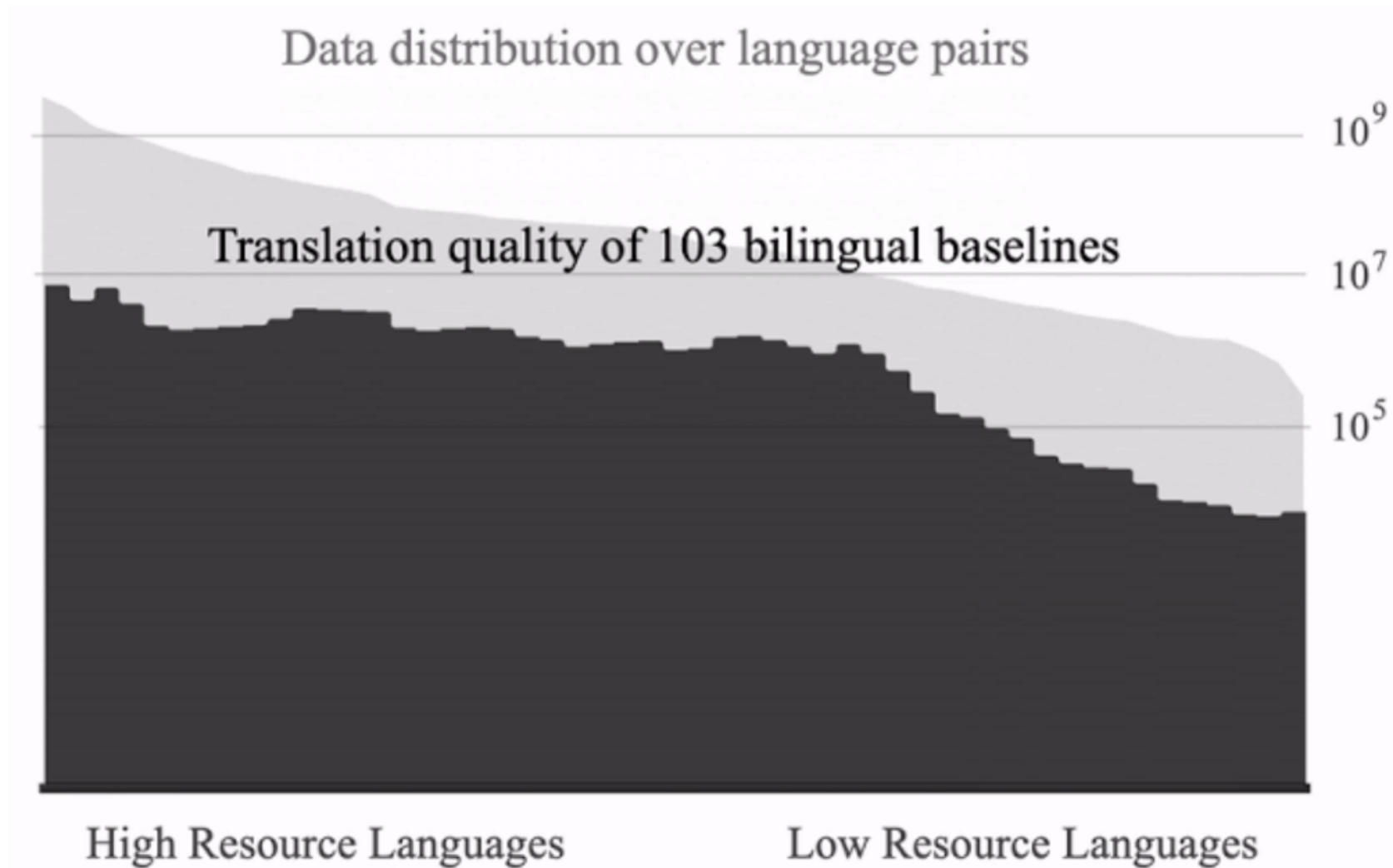
(source: Google)

Gains with Multilingual Training

- Massively multilingual with 50 billion parameters
- Massively multilingual with 6 billion parameters
- Massively multilingual with 400 million parameters

+15 BLEU

+10 BLEU

+5 BLEU

Bilingual Baselines →

-5 BLEU

High Resource Languages   Low Resource Languages
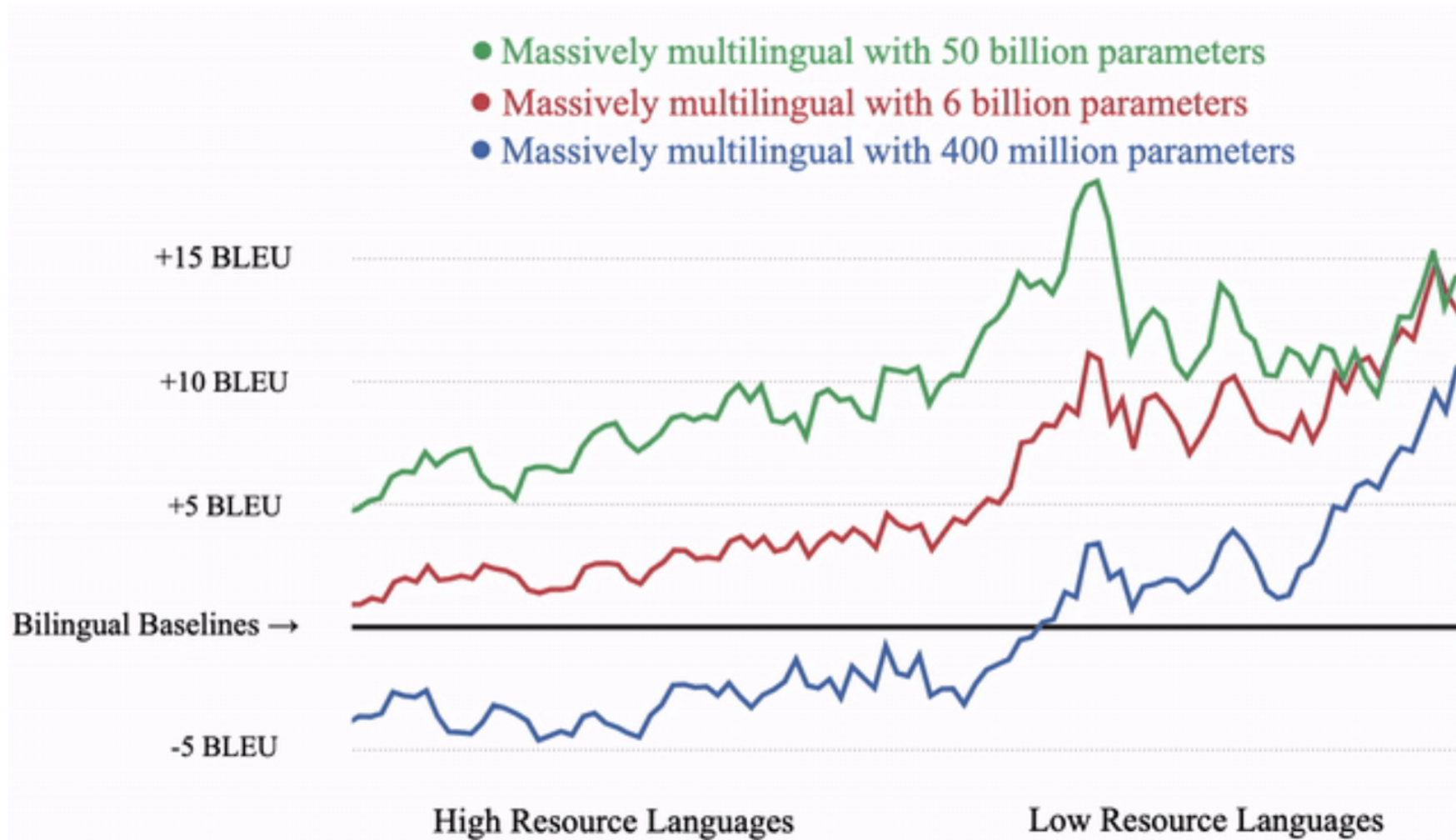
(source: Google)

Romanization

ፊሊፒንሲ		filipenesi		fil ipe nes i
الفلبين		alflbyn		alf lb yn
ফিলিপাইন		Philipaaina		Phil ipa aina
فيليپين		filipin		fil ipin
Fülöp-szigetek	Maximize shared orthography	Fueloep-szigetek	Maximize shared vocabulary	Fue lo ep - szi gete k
Filipina		Filipina		Filipin a
Филиппины		Filippiny		Fili ppi ny
Filibiin		Filibiin		Fili bi in
Filipinas	uroman	Filipinas	BPE	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)

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

multi-task training

- Our translation models: generic sequence-to-sequence models

- Same model used for many other tasks

  - sentiment detection
  - grammar correction
  - semantic inference
  - summarization
  - question answering
  - speech recognition

- For all these tasks, we need to learn basic properties of language

  - word embeddings
  - contextualize word representations in encoder
  - language model aspects of decoder

- Why re-invent the wheel each time?

- Train model on several tasks

- Maybe shared and task-specific components

- System learns general facts about language

  – informed by many different tasks

  – useful for many different tasks

Pre-Training Word Embeddings

- Let us keep it simple...

- Neural machine translation models use word embeddings

    – encoding of input words
    – encoding of output words

- Word embeddings can be trained on vast amounts of monolingual data

⇒ pre-train word embeddings and initialize model with them

- Not very successful so far

    – monolingual word embeddings trained on language model objectives
    – for machine translation, different similarity aspects may matter more
    – e.g., *teacher* and *teaching* similar in MT, not in LM

Pre-Training the Encoder and Decoder

- Pre-training other components of the translation model

- Decoder

  - language model, informed by input context
  - pre-train as language model on monolingual data
  - input context vector set to zero

- Encoder

  - also structures like a language model

    (however, not optimized to predict following words)
  - pre-train as language model on monolingual data

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

Multi-Task Training

- Multiple end-to-end tasks that share common aspects
  - need to encode an input word sequence
  - produce an output word sequence

- May have very different input/output
  - sentiment detection: output is sentiment value
  - part-of-speech tagging: output is tag sequence
  - syntactic parsing: output is recursive parse structure (may be linearized)
  - semantic parsing: output is logical form, database query, or AMR
  - grammar correction: input is error-prone text
  - question answering: needs to be also informed by knowledge base
  - speech recognition: input is sequence of acoustic features

- Input and output in the same language, may be mostly copied
  - grammar correction, automatic post-editing
  - question answering, semantic inference

Multi-Task Training

- Train a single model for all tasks

- Positive results with joint training of

  - part-of-speech tagging
  - named entity recognition
  - syntactic parsing
  - semantic analysis.

- Tasks may share just some components