Multilingual Machine Translation

and Machine Translation for Multilinguality

Ondřej Bojar

May 22, 2020

EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Outline

Part 1: Tools from MT for Exploring Multilinguality.

- Parallel and multi-lingual corpora.
- Sentence alignment.
- Word alignment.

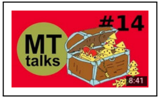
Part 2: Exploiting Multilinguality for MT.

- Motivation for more than two languages in MT.
- Interesting configurations.
- Dedicated architectures vs. simple data mixing.
- Interlingua?

Embedded slides by Rico Sennrich and Adam Lopez.

Supplementary Materials

Videolectures & Wiki:

http://mttalks.ufal.ms.mff.cuni.cz/ 

NPFL087 Slides and Lectures:

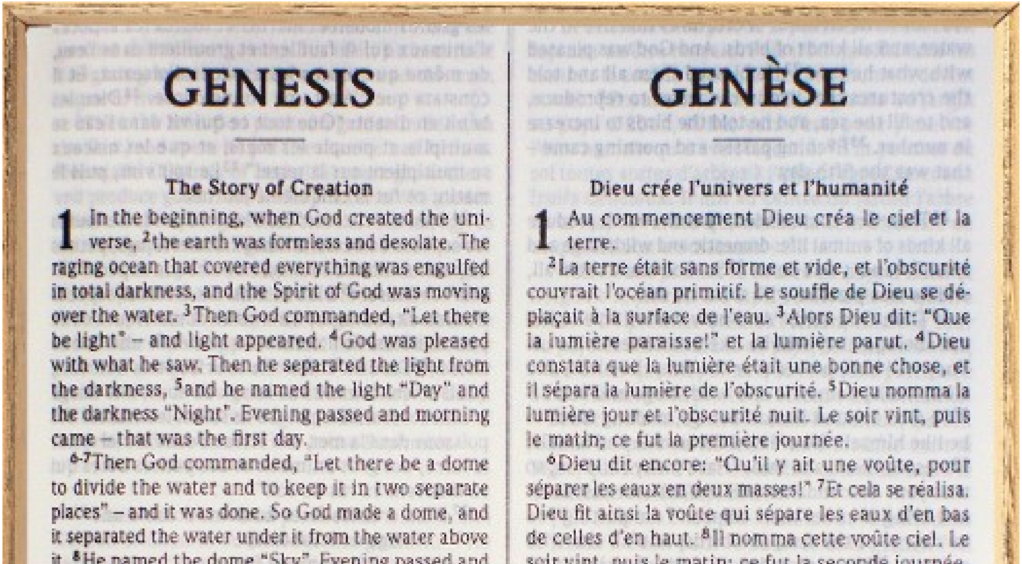
http://ufal.mff.cuni.cz/courses/npfl087

Books:

- Ondřej Bojar: Čeština a strojový překlad. ÚFAL, 2012.
- Philipp Koehn: Statistical Machine Translation. Cambridge University Press, 2009.
  With some slides: http://statmt.org/book/
  NMT: https://arxiv.org/pdf/1709.07809.pdf

Tools from MT for Exploring Multilinguality

A Classical Parallel Corpus



GENESIS

The Story of Creation

1 In the beginning, when God created the uni-
verse, 2 the earth was formless and desolate. The
raging ocean that covered everything was engulfed
in total darkness, and the Spirit of God was moving
over the water. 3 Then God commanded, "Let there
be light" – and light appeared. 4 God was pleased
with what he saw. Then he separated the light from
the darkness, 5 and he named the light "Day" and
the darkness "Night". Evening passed and morning
came – that was the first day.
6-7 Then God commanded, "Let there be a dome
to divide the water and to keep it in two separate
places" – and it was done. So God made a dome, and
it separated the water under it from the water above
it. 8 He named the dome "Sky". Evening passed and

GENÈSE

Dieu crée l'univers et l'humanité

1 Au commencement Dieu créa le ciel et la
terre.
2 La terre était sans forme et vide, et l'obscurité
couvrait l'océan primitif. Le souffle de Dieu se dé-
plaçait à la surface de l'eau. 3 Alors Dieu dit: "Que
la lumière paraisse!" et la lumière parut. 4 Dieu
constata que la lumière était une bonne chose, et
il sépara la lumière de l'obscurité. 5 Dieu nomma la
lumière jour et l'obscurité nuit. Le soir vint, puis
le matin; ce fut la première journée.
6 Dieu dit encore: "Qu'il y ait une voûte, pour
séparer les eaux en deux masses!" 7 Et cela se réalisa.
Dieu fit ainsi la voûte qui sépare les eaux d'en bas
de celles d'en haut. 8 Il nomma cette voûte ciel. Le
soir vint, puis le matin; ce fut la seconde journée.

Another Classical One (1658)

Invitatio. Einleitung.

M. Veni, Puer! diſce Sapere.
P. Quid hoc eſt, *Sapere?*
M. Omnia,

L. Kom̃ her/ Knab! lerne Weißheit.
S. Was iſt das/ Weißheit?
L. Alles/

Parallel Corpora

- Web is an immense resource.
- People keep crawling it over and over:
  - Bitextor: Esplà-Gomis and Forcada (2010)
  - `http://paracrawl.eu/releases.html` (2018)
- Good sources of (multi-)parallel corpora:
  - Corpus OPUS: `http://opus.nlpl.eu/`
  - UN Corpus, various EU corpora (DGT-Acquis)...
  - WMT tasks data: `http://www.statmt.org/wmt20/`
  - University-specific corpora, e.g. UFAL released:
    - `http://ufal.mff.cuni.cz/czeng` (Czech-English)
    - `http://ufal.mff.cuni.cz/hindencorp` (Hindi-English), ..., Odia-English...
    - `http://ufal.mff.cuni.cz/umc/`
      (Czech, Russian, Urdu, with English)

Aligned Documents ⇝ Sentence Pairs

In my dream , there was a sycamore growing out of the ruins of the sacristy , and I was told that , if I dug at the roots of the sycamore , I would find a hidden treasure . But I ' m not so stupid as to cross an entire desert just because of a recurrent dream . " And they disappeared . The boy stood up shakily , and looked once more at the Pyramids . " It is I who dared to do so , " said the boy . This man looked exactly the same , except that now the roles were reversed . " It is I who dared to do so , " he

अपने सपने में मुझे एक गुलर का पेड़ दिखाई देता था और मुझे लगता था कि अगर मैं उस गुलर की जड़ें खोद डालूं तो मुझे छिपा हुआ खजाना मिल जाएगा । मगर मैं तुम्हारी तरह इतना बेवकूफ नहीं हूं कि महज बार – बार आने वाले एक सपने के कारण पूरे रेगिस्तान को पार करूं । वे लोग , उसके बाद वहां से चले गए । लड़का लड़खड़ाता हुआ किसी तरह खड़ा हो गया ।<s>एक बार फिर उसने पिरामिडों को देखा । “ यह जुर्रत मैंने की थी , ” लड़के ने कहा ।<s>उसे सेंटियागो मातामोरोस कीं वह प्रतिमा याद आई जिसमें वह घोड़े पर सवार था और उसके घोड़े के खुरों में कितने ही नास्तिक कुचले हुए पड़े थे । यह घुड़सवार भी बिलकुल वैसा ही था । यह बात और थी कि इनके किरदार बदले हुए थे । “ मैंने ही ऐसा करने का साहस किया था , ” लड़के ने दोहराया और अपनी गर्दन तलवार का वार सहने के लिए झुका दी । ‘ जिंदगी ने भी हमेशा मेरे साथ अच्छा बर्ताव किया । ’

Aligned Documents ⇝ Sentence Pairs

In my dream , there was a sycamore growing out of the ruins of the sacristy , and I was told that , if I dug at the roots of the sycamore , I would find a hidden treasure .	अपने सपने में मुझे एक गुलर का पेड़ दिखाई देता था और मुझे लगता था कि अगर मैं उस गुलर की जड़ें खोद डालूं तो मुझे छिपा हुआ खजाना मिल जाएगा ।
But I ’ m not so stupid as to cross an entire desert just because of a recurrent dream . "	मगर मैं तुम्हारी तरह इतना बेवकूफ नहीं हूं कि महज बार - बार आने वाले एक सपने के कारण पूरे रेगिस्तान को पार करूं ।
And they disappeared .	वे लोग , उसके बाद वहां से चले गए ।
The boy stood up shakily , and looked once more at the Pyramids .	लड़का लड़खड़ाता हुआ किसी तरह खड़ा हो गया । एक बार फिर उसने पिरामिडों को देखा ।
	“ यह जुर्रत मैंने की थी , ” लड़के ने कहा । उसे सेंटियागो मातामोरोस कीं वह प्रतिमा याद आई जिसमें वह घोड़े पर सवार था और उसके घोड़े के खुरों में कितने ही नास्तिक कुचले हुए पड़े थे ।
" It is I who dared to do so , " said the boy .	यह घुड़सवार भी बिलकुल वैसा ही था ।
This man looked exactly the same , except that now the roles were reversed .	यह बात और थी कि इनके किरदार बदले हुए थे ।
" It is I who dared to do so , " he repeated , and he lowered his head to receive a blow from the sword .	“ मैंने ही ऐसा करने का साहस किया था , ” लड़के ने दोहराया और अपनी गर्दन तलवार का वार सहने के लिए झुका दी ।
" Life was good to me , " the man said .	‘ जिंदगी ने भी हमेशा मेरे साथ अच्छा बर्ताव किया । ’
" When you appeared in my dream , I felt that all my efforts had been rewarded , because my son ’ s poems will be read by men for generations to come .	उस आदमी ने कहा , ‘ जब आप मेरे सपने में आए थे , तो मुझे लगा कि मैंने अपने कर्मों का पुरस्कार पा लिया …
	मेरे लिए इससे बढ़कर और क्या बात होती कि मेरे बेटे की कविताएं युग - युगों तक पदी जाएं ।
I don ’ t want anything for myself .	नहीं , मुझे अपने लिए कुछ नहीं चाहिए ।
But any father would be proud of the fame achieved by one whom he had cared for as a child , and educated as he grew up .	कोई भी बाप उस इंसान की शोहरत सुनकर फूला नहीं समाएगा जिसे उसने अपनी गोद में खिलाया , पढ़ाया - लिखाया और पाल - पोसकर बड़ा किया हो ।
" We ’ re two very different things . "	“ हम दो अलग - अलग चीजें हैं । ”
" That ’ s not true , " the boy said .	“ यह सही नहीं है । ” लड़के ने कहा ,
" I learned the alchemist ’ s secrets in my travels .	“ यात्रा के दौरान मैंने कीमियागर के रहस्यों को जाना है ।
I have inside me the winds , the deserts , the oceans , the stars , and everything created in the universe .	मेरे ही भीतर सब छिपा है — हवा , रेगिस्तान , समुद्र , तारे और वह सब कुछ जो ब्रह्माण्ड ने सर्जित किया है ।
We were all made by the same hand , and we have the same soul .	हम सबको उसी हाथ ने बनाया और हम सबकी आत्मा भी एक ही है ।
You ’ ll learn to love the desert , and you ’ ll get to know every one of the fifty thousand palms .	तुम्हें रेगिस्तान से प्यार करना आ जाएगा और उन पचास हजार खजूर के पेड़ों में तुम एक - एक को पहचानने लगोगे ।
You ’ ll watch them as they grow , demonstrating how the world is always changing .	उन्हें बढ़ता हुआ देखकर तुम अनुभव करोगे कि कैसे हर क्षण दुनिया बदलती रहती है ।
And you ’ ll get better and better at understanding omens , because the desert is the best teacher there is .	तुम शकुन पहचानने में बेहतर से बेहतर बनते जाओगे चूंकि इस मामले में रेगिस्तान से बढ़कर कोई अच्छा गुरु नहीं है ।
" Sometime during the second year , you ’ ll remember about the treasure .	“ फिर , किसी वक्त , दूसरे साल के दौरान तुम्हें खजाने की याद सताएगी ।
The omens will begin insistently to speak of it , and you ’ ll try to ignore them .	शकुन फौरन तुम्हें उसके बारे में बताना शुरू कर देंगे , मगर तुम उन्हें अनदेखा करना चाहोगे ।
But you know that I ’ m not going to go to Mecca . Just as you know that you ’ re not going to buy your sheep . "	तुम अच्छी तरह से जानते हो , कि मैं मक्का नहीं जाने वाला हूं ठीक उसी तरह जैसे कि तुम कोई भेड़ - वेड़ नहीं खरीदने वाले हो ! ”
" Who told you that ? " asked the boy , startled .	“ आपसे ऐसा किसने कहा ? ” लड़के को आश्चर्य हुआ ।
" Maktub " said the old crystal merchant .	“ मकतूब ! ” क्रिस्टल - व्यापारी ने कहा ,
And he gave the boy his blessing .	कुछ पल खामोश रह कर , उसने लड़के को भरपूर आशीर्वाद दिया ।
The boy went to his room and packed his belongings .	कमरे में जाकर लड़के ने अपना सामान बांधा ।
They filled three sacks .	तीन बोरे भर गए ।
As he was leaving , he saw , in the corner of the room , his old shepherd ’ s pouch .	बाहर जाते हुए उसने कमरे के एक कोने में , अपनी पुरानी थैली देखी ।
" I want to see the greatness of Allah , " the chief said , with respect .	“ मैं अल्लाह की महानता देखना चाहता हूं । ” बड़े आदर के साथ मुखिया ने कहा ।
" I want to see how a man turns himself into the wind . "	“ मैं देखना चाहता हूं कि कैसे कोई आदमी खुद को हवा में बदलता है । ”
But he made a mental note of the names of the two men who had expressed their fear .	मगर उसने अपने मन में उन दो सेनापतियों के नाम याद कर लिए जिन्होंने डर का इजहांर किया था ।

Sentence Alignment

Goal: Given a text in two languages, align sentences.

Assume: Sentences hardly ever reordered.

- Classical algorithm: Gale and Church (1993).
    - Based on similar character length of aligned sentences, no words examined.
    - Dynamic-programming search for the best alignment.
    - Allows 0 to 2 sentences in a group: 0-1, 1-0, 1-1, 2-1, 1-2, 2-2.
- Several algorithms for English-Czech evaluated by Rosen (2005).
    - Nearly perfect alignment possible by a combination of aligners.
- The "standard tool": Hunalign (Varga et al., 2005).
    - LF Aligner has even a user interface for correcting alignments.
- Another option: Gargantua (Braune and Fraser, 2010).

MT Talk #7

http://mttalks.ufal.ms.mff.cuni.cz/index.php?title=Sentence_Alignment

Word Alignment

Goal: Given a sentence in two languages, align words (tokens).
State of the art: GIZA++ (Och and Ney, 2000):

- Unsupervised, only sentence-parallel texts needed.
- Word alignments formally restricted to a function:

$$\text{src token} \mapsto \text{tgt token or NULL}$$

- A cascade of models refining the probability distribution:
  - IBM1: only lexical probabilities: $P(\textit{kočka} = \textit{cat})$
  - IBM3: adds fertility: 1 word generates several others
  - IBM4/HMM: to account for relative reordering
- Only many-to-one links created $\Rightarrow$ used twice, in both directions.

IBM Model 1

"Model" $=$ word-for-word translation dictionary, $P(\textit{kočka} \mid \textit{cat})$

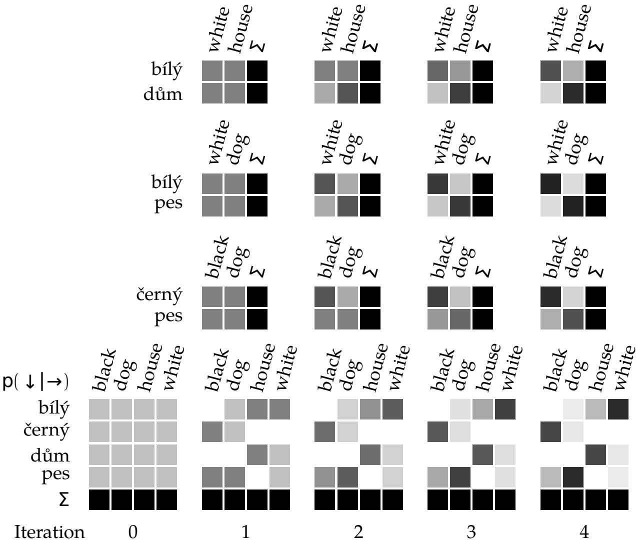
$=$ "Lexical probabilities" only, positions of words disregarded.

Probabilities estimated using Expectation-Maximization Loop:

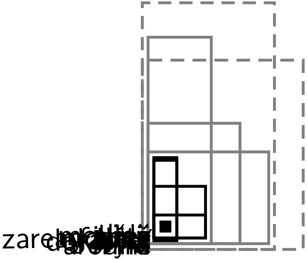
1. Start by assuming any word can be translated as any word.
   …i.e. a dictionary with flat probabilities.
2. (Expectation) Draw alignment links in sentences based on dict.
   …this will be flat, every word with every word, in the first loop.
3. (Maximization) Set probabilities in the dict based on cooc. counts.
4. Go to Step 2.

MT Talk #8: http://mttalks.ufal.ms.mff.cuni.cz/index.php?title=Word_Alignment

EM Loop in IBM1



Phrase-Based MT Overview



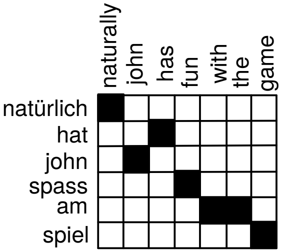
This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
...	=	...
This time around, they 're moving	=	Nyní zareagovaly
even faster	=	dokonce ještě rychle
...	=	...

Phrase-based MT: choose such segmentation of input string and such phrase "replacements" to make the output sequence "coherent" (3-grams most probable).

Extracting Linguistic Patterns (1/3)

Phrase extraction for standard phrase-based MT:

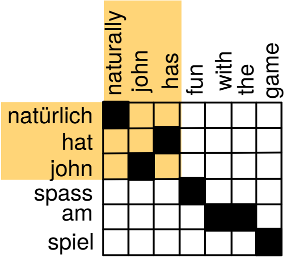
1. Run sentence and word alignment,



Extracting Linguistic Patterns (2/3)

Phrase extraction for standard phrase-based MT:

1. Run sentence and word alignment,
2. Extract all phrases consistent with word alignment.

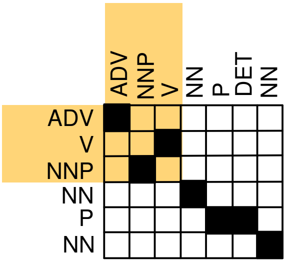


$\Rightarrow$ Extracted: natürlich hat john $\rightarrow$ naturally john has

Extracting Linguistic Patterns (3/3)

Now reused for extracting some other linguistic correspondences:

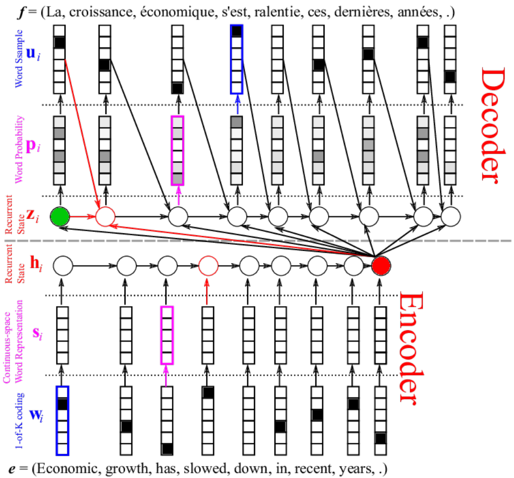
1. Run sentence and word alignment,
2. Extract same phrases, but e.g. POS tags, not word forms.



$\Rightarrow$ Extracted: ADV V NNP $\rightarrow$ ADV NNP V

Exploiting Multilinguality for MT

Neural MT: Encoder-Decoder



https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-2/

Why More than Two Languages?

- Help in low-resource settings.
  - Words, morphemes or syntactic patterns common to more languages.
  - Learning can reuse patterns seen in another dataset.
- Improve translation quality.
  - Words are ambiguous, the third language can disambiguate.
- Truly multi-lingual environments.
  - United Nations: 6 languages.
  - EU official languages: 24.
  - EUROSAI official languages: 43.
  - INTOSAI official languages...

Multilingual MT Configurations

- Pivot translation (Cascading).
- Multi-lingual source (also called multi-way).
- Multi-lingual multi-source.
- Multi-lingual target.
- Multi-lingual multi-target.
- Both sides multi-lingual.
- (Both sides multi-lingual, multi-source, multi-target. ;-)
- Zero-shot training.
    - i.e. translating an unseen pair when both the source and target langs were covered in the training data in other pairs.
- "Beyond zero-shot" is translating from an unseen language.

ELITR Multi-Target and Multi-Source MT

- Multi-Target focus: Efficiency
  - Decrease hardware resources compared using many separate models.
- Multi-Source focus: Resolving ambiguity thanks to existing translation
  - E.g. Translating German “Schloss” to French is easier if we can feed in the English translation (“castle” or “lock”).
- Training on: Multi-parallel or bi-parallel multilingual corpora.

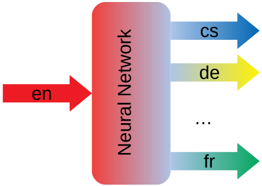


Figure 1: Multi-Target MT

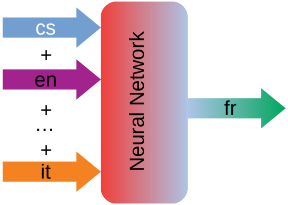


Figure 2: Multi-Source MT

ELITR Y3 Goal: Flexible Multi-Lingual MT

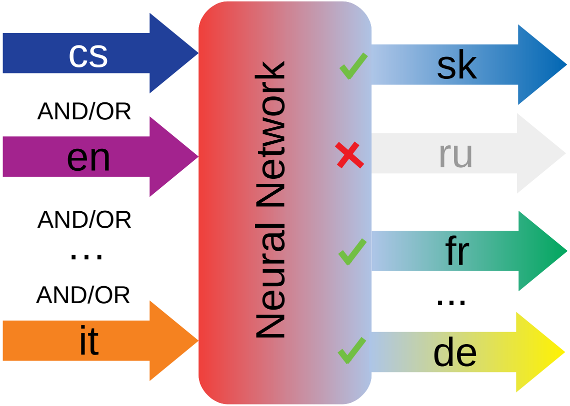


Figure 3: Flexible multilingual MT

Strategies for NMT

- Simple data mixing.
  - Multilingual models.
  - Pre-training / Transfer learning.
- Dedicated architectures.

Simple Data Mixing

... simply feed in various language pairs.

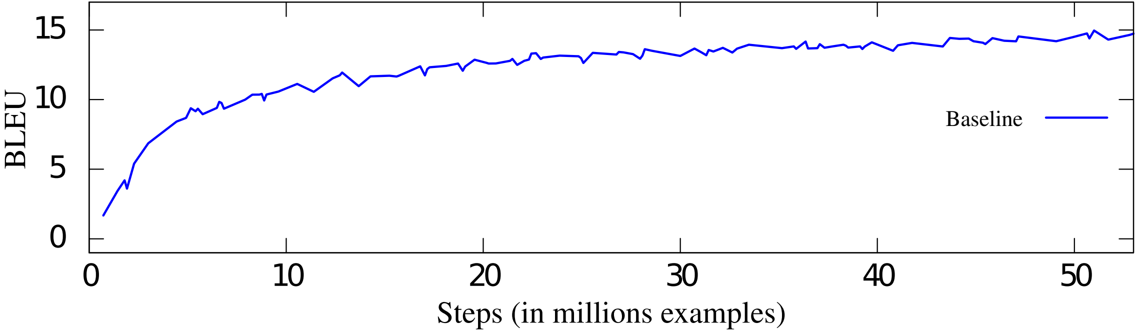
Source Sent 1 (De)	**2en** versetzen Sie sich mal in meine Lage !
Target Sent 1 (En)	put yourselves in my position .
Source Sent 2 (En)	**2nl** I flew on Air Force Two for eight years .
Target Sent 2 (Nl)	ik heb acht jaar lang met de Air Force Two gevlogen .

- The model of the same size will learn both pairs.
- Hopefully benefiting from various similarities.
- Risk of catastrophic forgetting.

See Johnson et al. (2016) or Ha et al. (2017).

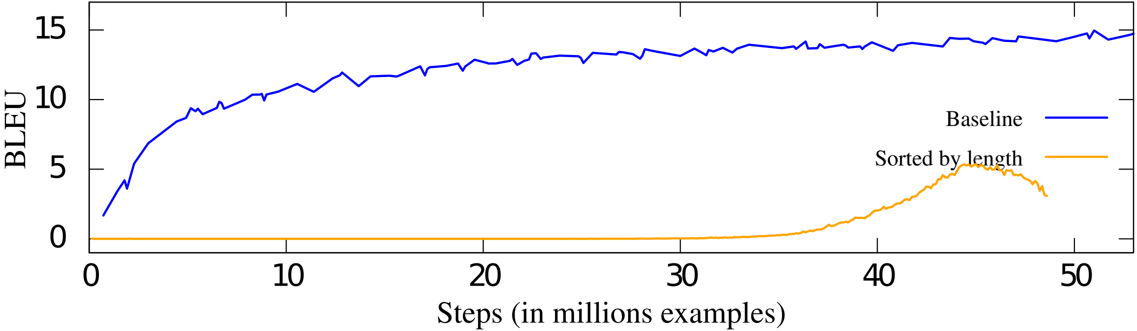
Interlude: Catastrophic Forgetting

- Kocmi and Bojar (2017) explore curriculum learning:
  - Start with simpler sentences first, add complex ones later.



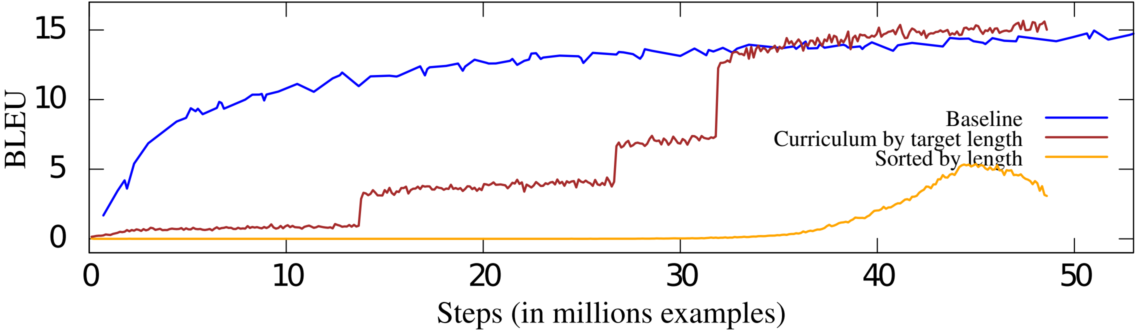
Interlude: Catastrophic Forgetting

- Kocmi and Bojar (2017) explore curriculum learning:
  - Start with simpler sentences first, add complex ones later.
- When "simpler" means "shorter":



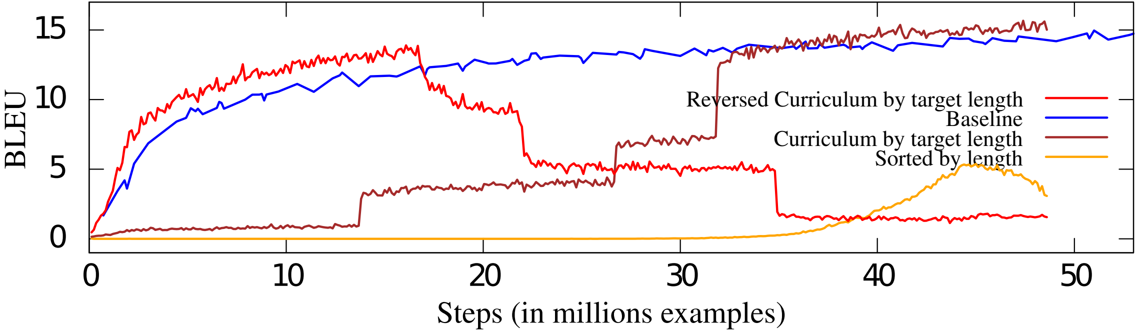
Interlude: Catastrophic Forgetting

- Kocmi and Bojar (2017) explore curriculum learning:
  - Start with simpler sentences first, add complex ones later.
- When "simpler" means "shorter":
  - Clear jumps in score as bins of longer sentences are allowed.

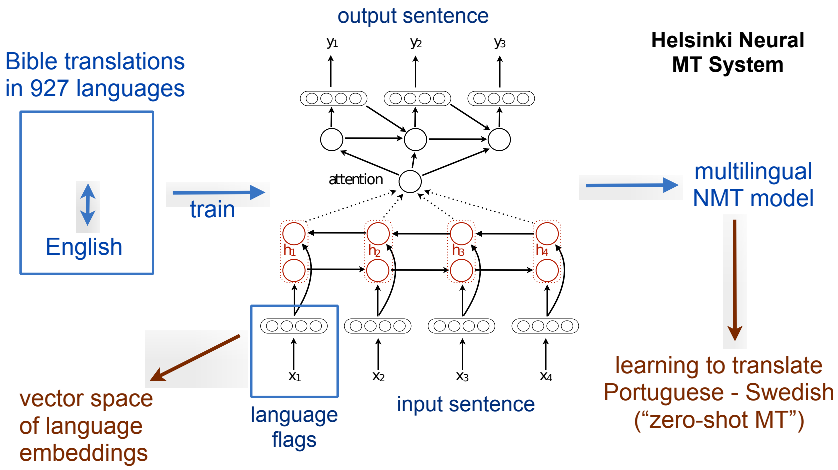


Interlude: Catastrophic Forgetting

- Kocmi and Bojar (2017) explore curriculum learning:
  - Start with simpler sentences first, add complex ones later.
- When “simpler” means “shorter”:
  - Clear jumps in score as bins of longer sentences are allowed.
  - Reversed curriculum <u>unlearns</u> to produce long sentences.

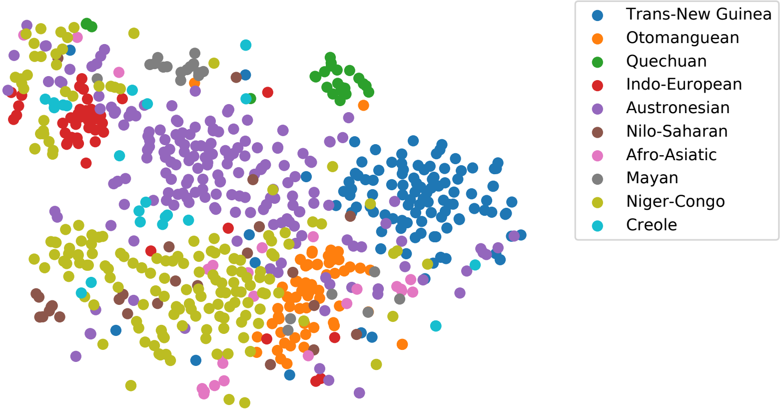


“Language Embeddings” from 927 Bibles



Tiedemann (2018)

"Language Embeddings" from 927 Bibles

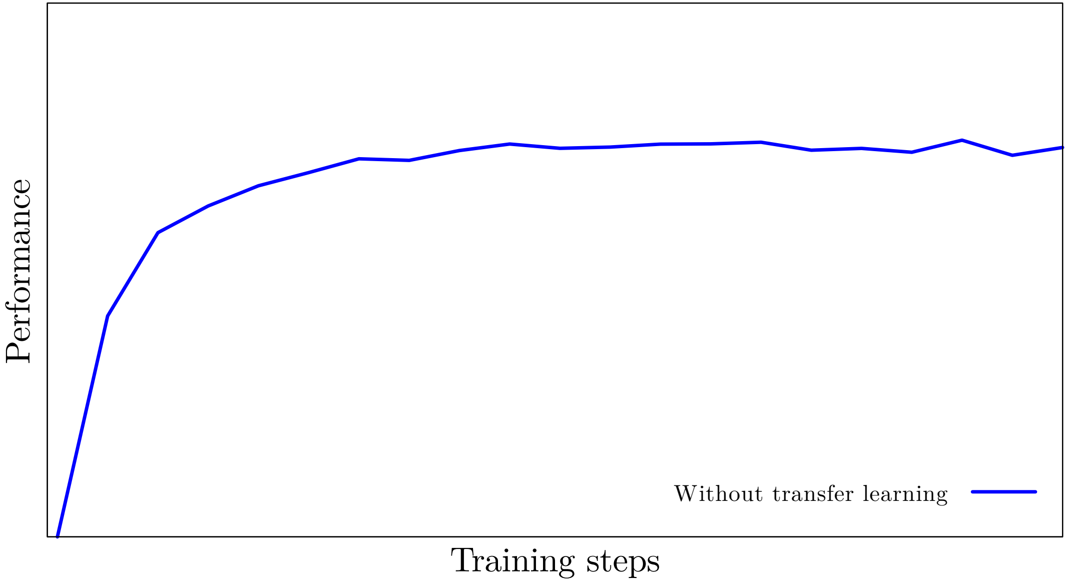


t-SNE of the language-embedding vectors, colored by language family.

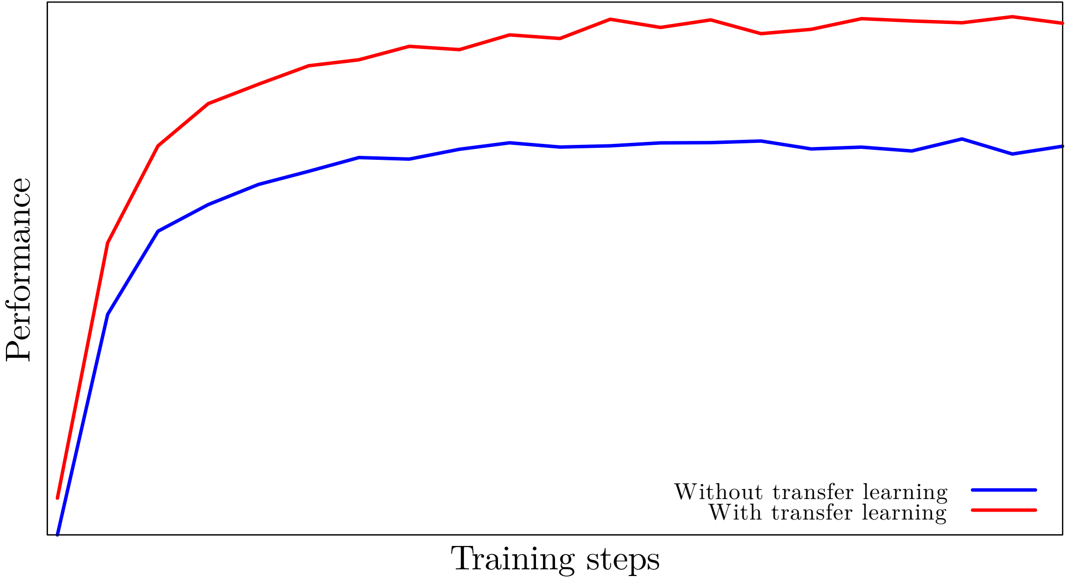
Exploiting Multilinguality for MT

Transfer Learning

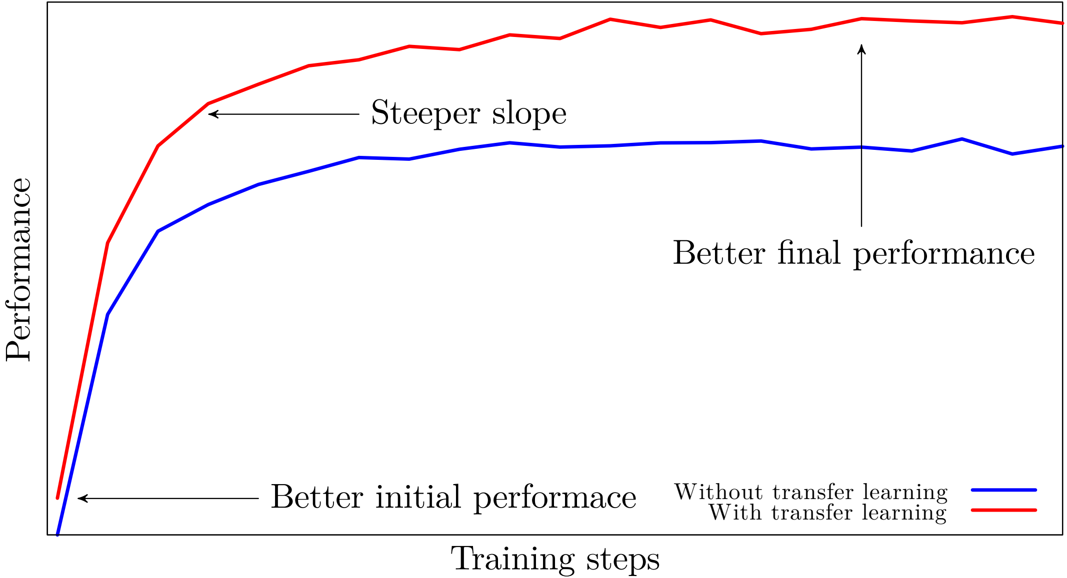
Motivation for NN Transfer Learning



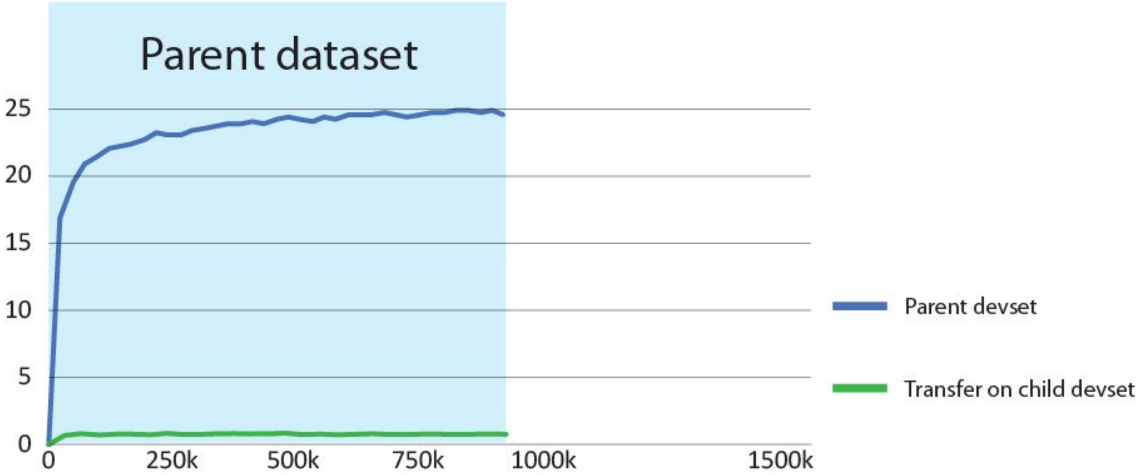
Motivation for NN Transfer Learning



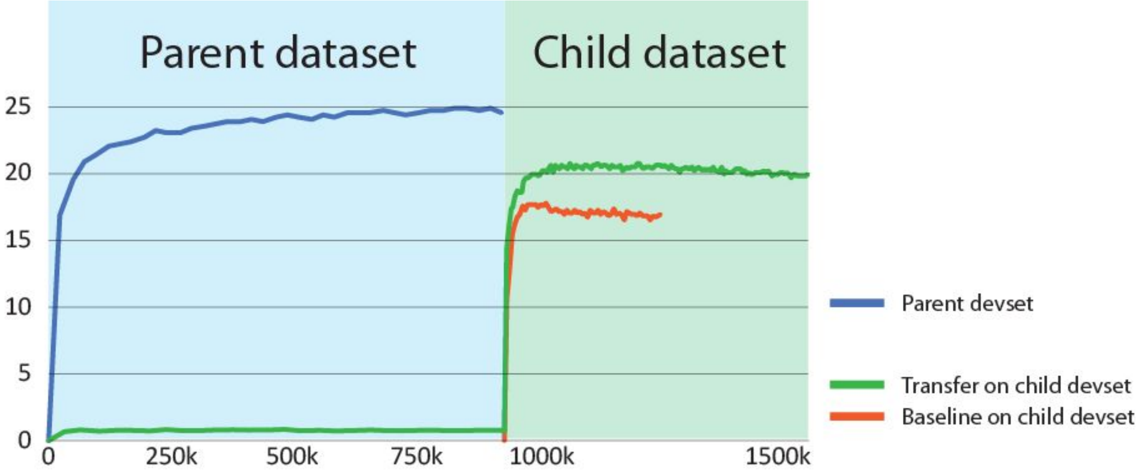
Motivation for NN Transfer Learning



Steps of Transfer Learning



Steps of Transfer Learning



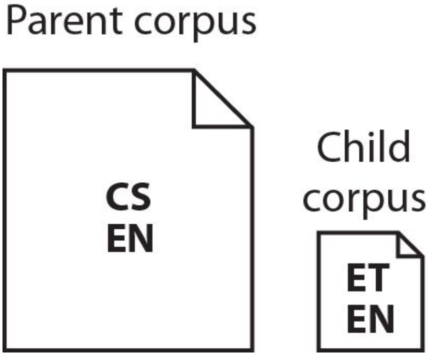
Trivial Transfer Learning

- Early works (Zoph et al., 2016; Nguyen and Chiang, 2017) target one common language (English).
- Kocmi and Bojar (2018) try even unrelated languages.

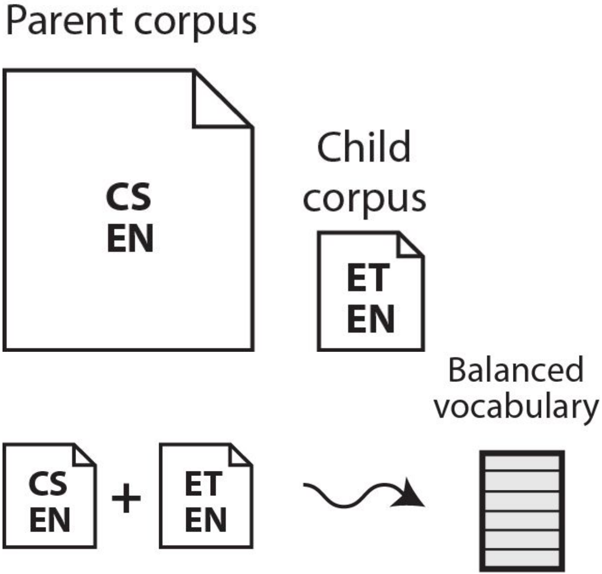
The trivial procedure:

- Train on one pair (“parent”), switch corpus to another (“child”).
- The only requirement: joint subword units across all langs.

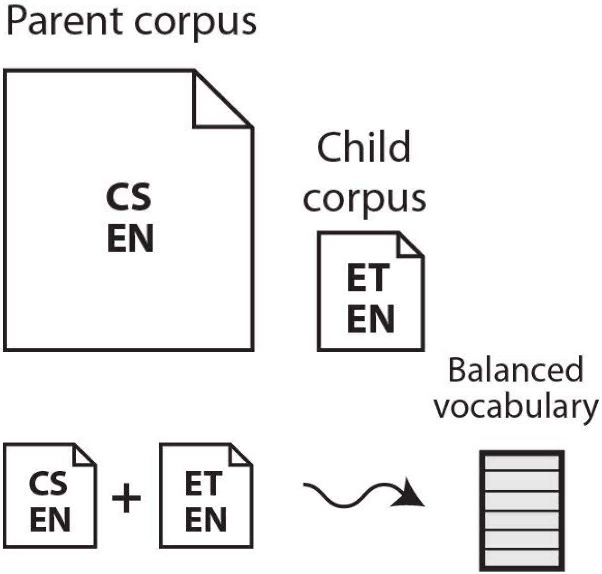
Getting Balanced Vocabulary



Getting Balanced Vocabulary



Getting Balanced Vocabulary




```
the_
že_
ying_
staying_
pra
pracovat_
...
```

English on Same Side

Child model: Slovak

Parent model	**Corpus size difference**	**Direction**	**Baseline (BLEU)**	**Transfer (BLEU)**	**Δ (BLEU)**
Czech	9x	from English	16.13	17.75	1.62 *
Czech	9x	to English	19.19	22.42	3.23 *

* statistically significant

English on Same Side

Child model: Slovak

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Czech	9x	from English	16.13	17.75	1.62 *
Czech	9x	to English	19.19	22.42	3.23 *

* statistically significant

English on Same Side

Child model: Slovak

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Czech	9x	from English	16.13	17.75	1.62 *
Czech	9x	to English	19.19	22.42	3.23 *

Child model: Estonian

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Finnish	3.5x	from English	17.03	19.74	2.71 *
Russian	16x	from English	17.03	20.09	3.06 *
Czech	50x	from English	17.03	20.41	3.38 *
Finnish	3.5x	to English	21.74	24.18	2.44 *
Russian	16x	to English	21.74	23.54	1.80 *

* statistically significant

English on Same Side

Child model: Slovak

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Czech	9x	from English	16.13	17.75	1.62 *
Czech	9x	to English	19.19	22.42	3.23 *

Child model: Estonian

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Finnish	3.5x	from English	17.03	19.74	2.71 *
Russian	16x	from English	17.03	20.09	3.06 *
Czech	50x	from English	17.03	20.41	3.38 *
Finnish	3.5x	to English	21.74	24.18	2.44 *
Russian	16x	to English	21.74	23.54	1.80 *

* statistically significant

English on Same Side

Child model: Slovak

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Related	9x	from English	16.13	17.75	1.62 *
Related	9x	to English	19.19	22.42	3.23 *

Child model: Estonian

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Related	3.5x	from English	17.03	19.74	2.71 *
Cyrillic	16x	from English	17.03	20.09	3.06 *
Biggest	50x	from English	17.03	20.41	3.38 *
Related	3.5x	to English	21.74	24.18	2.44 *
Cyrillic	16x	to English	21.74	23.54	1.80 *

* statistically significant

English on Same Side, Parent Low-Resource

Child model: Finnish

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Estonian	0.3x	from English	19.50	20.07	0.57 *
Estonian	0.3x	to English	24.40	23.95	-0.45

Child model: Czech

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Slovak	0.1x	from English	23.48	22.99	-0.49 *
Slovak	0.1x	to English	29.61	28.20	-1.41 *

English on Same Side, Parent Low-Resource

Child model: Finnish

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Estonian	0.3x	from English	19.50	20.07	0.57 *
Estonian	0.3x	to English	24.40	23.95	-0.45

Child model: Czech

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Slovak	0.1x	from English	23.48	22.99	-0.49 *
Slovak	0.1x	to English	29.61	28.20	-1.41 *

English on the Other Side

Parent model	Child model	Corpus size amplification	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)	Parent Aligned Δ
EN - Finnish	Estonian - EN	3.5x	21.74	22.75	1.01 *	2.44 *
EN - Russian	Estonian - EN	16x	21.74	23.12	1.38 *	1.80 *
EN - Czech	Estonian - EN	50x	21.74	22.80	1.06 *	
Finnish - EN	EN - Estonian	3.5x	17.03	18.19	1.16 *	2.71 *
Russian - EN	EN - Estonian	16x	17.03	18.16	1.13 *	3.06 *

No Language in Common

Child model: Estonian to English

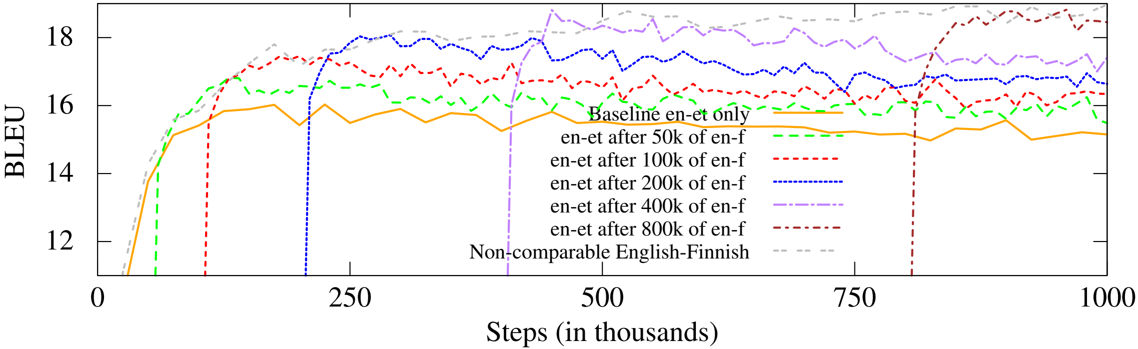
Parent model	Corpus size amplification	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Arabic - Russian	12x	21.74	22.23	0.49
Spanish - French	12x	21.74	22.24	0.50 *
Spanish - Russian	12x	21.74	22.52	0.78 *
French - Russian	12x	21.74	22.40	0.66 *

No Language in Common

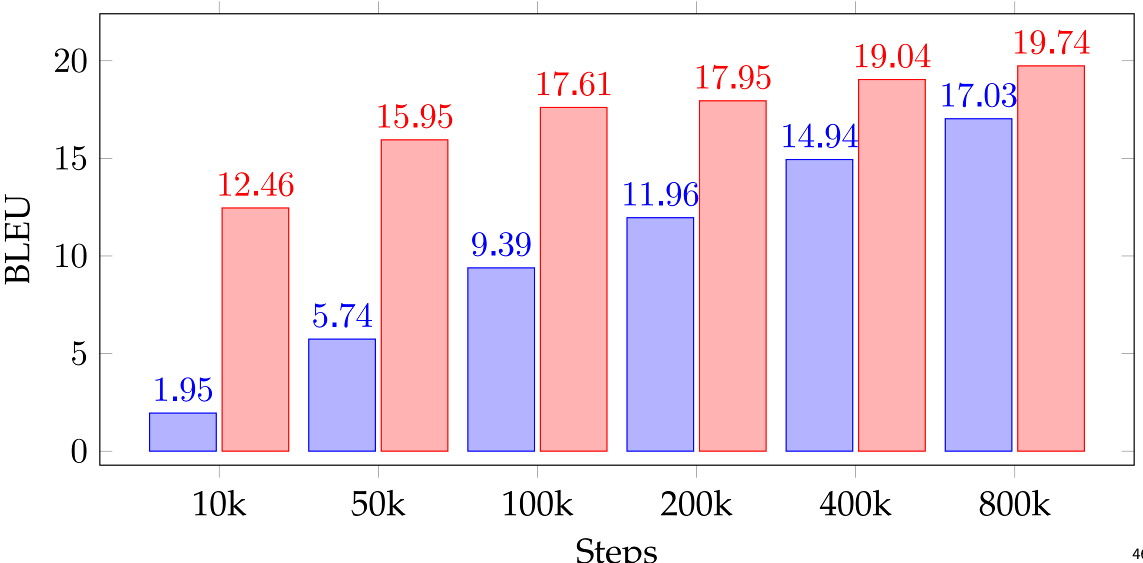
Child model: Estonian to English

Parent model	Corpus size amplification	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Arabic Cyrillic	2x	21.74	22.23	0.49
Spanish - French	12x	21.74	22.24	0.50 *
Spanish - F Cyrillic	2x	21.74	22.52	0.78 *
French - R Cyrillic	2x	21.74	22.40	0.66 *

The Better the Parent, the Better the Child



The Lesser the Child, the Bigger the Gain



Why it Helps? Not Really Vocabulary (1/2)

	Length	BLEU Components	BP
Base ENET	35326	48.1/21.3/11.3/6.4	0.979
ENRU+ENET	35979	51.0/24.2/13.5/8.0	0.998
ENCS+ENET	35921	51.7/24.6/13.7/8.1	0.996

(The reference length in the matching tokenization was 36062.)

- Child models produce longer outputs $\Rightarrow$ lower brevity penalty.
- But n-gram precisions also better.

1-gram present in	ENRU+ENET	ENCS+ENET
Child, Base, Ref	15902 (44.2 %)	15924 (44.3 %)
Child only	9635 (26.8 %)	9485 (26.4 %)
Child, Base	7209 (20.0 %)	7034 (19.6 %)
Child, Ref	**3233 (9.0 %)**	**3478 (9.7 %)**
Total	35979 (100.0 %)	35921 (100.0 %)

- The 3k better toks are regular ET words, not NEs or numbers.

Why it Helps? Not Really Vocabulary (2/2)

Why it Helps? Sentence Lengths Somewhat

	Parent		Child	
Sentence lengths	BLEU	Avg. words	BLEU	Avg. words
1-10 words	8.57	10.9	16.57	15.3
10-20 words	16.21	15.4	17.48	15.3
20-40 words	12.59	21.9	17.99	15.3
40-60 words	5.76	35.5	16.80	15.5
1-60 words	22.30	15.3	19.15	15.4

Why it Helps? Sentence Lengths Somewhat

	Parent		Child	
Sentence lengths	BLEU	Avg. words	BLEU	Avg. words
1-10 words	8.57	10.9	16.57	15.3
10-20 words	16.21	15.4	17.48	15.3
20-40 words	12.59	21.9	17.99	15.3
40-60 words	5.76	35.5	16.80	15.5
1-60 words	22.30	15.3	19.15	15.4

Why it Helps? Sentence Lengths Somewhat

	Parent		Child	
Sentence lengths	BLEU	Avg. words	BLEU	Avg. words
1-10 words	8.57	10.9	16.57	15.3
10-20 words	16.21	15.4	17.48	15.3
20-40 words	12.59	21.9	17.99	15.3
40-60 words	5.76	35.5	16.80	15.5
1-60 words	22.30	15.3	19.15	15.4

Exploiting Multilinguality for MT

Dedicated Architectures

Multi-source translation

Quite an old idea (e.g. Och & Ney 2001)

Table 4: Absolute improvements in WER combining two languages using method MAX compared with the best WER obtained by any of the two languages.

	fr	pt	es	it	sv	da	nl
fr	0.0	1.5	1.2	0.5	2.7	1.9	0.8
pt		0.0	2.2	2.1	4.0	3.4	1.3
es			0.0	2.4	3.9	2.6	1.7
it				0.0	3.5	3.2	1.6
sv					0.0	2.7	1.7
da						0.0	4.3
nl							0.0

Table 5: Absolute improvements in WER combining two languages using method PROD compared with the best WER obtained by any of the two languages.

	fr	pt	es	it	sv	da	nl
fr	0.0	0.8	0.1	0.4	1.0	0.8	-0.2
pt		0.0	2.6	2.1	2.6	2.8	-0.1
es			0.0	2.4	3.4	3.7	1.1
it				0.0	1.9	3.0	0.3
sv					0.0	1.8	0.5
da						0.0	1.5
nl							0.0

Table 6: Language combination using method MAX.

languages	WER	PER
fr	55.3	45.3
fr+sv	52.6	43.7
fr+sv+es	**52.0**	**43.2**
fr+sv+es+pt	52.3	43.6
fr+sv+es+pt+it	52.7	44.0
fr+sv+es+pt+it+da	52.5	43.9

Table 7: Language combination using method PROD.

languages	WER	PER
fr	55.3	45.3
fr+sv	54.3	44.5
fr+sv+es	51.0	41.4
fr+sv+es+pt	50.2	40.2
fr+sv+es+pt+it	49.8	39.8
fr+sv+es+pt+it+da	**48.8**	**39.1**

Multi-source translation

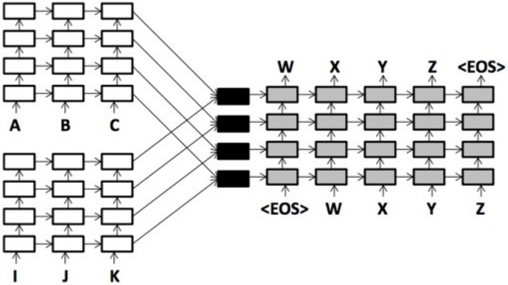
- Assorted techniques to do this in IBM-style or phrase-based MT.
- Difficult to model directly due to independence assumptions of these models.
- Usually done as a kind of system combination (merging the output of two MT systems).
- But this introduces other problems, e.g. decoding.
- Fundamentally, it's interpolation of conditional LMs.

Direct multi-source

Zoph & Knight 2016

- Directly learns and uses *p(English|French,German)*

- For attention: two context vectors (uses p-local attention of Luong, et al, but could use other methods).



Multi-way MT

Firat et al. 2016 (two papers)

- Assume only many bilingual parallel corpora.
- For N languages: learn N encoders and N decoders.
- But what about attention?

Multi-way MT

Firat et al. 2016 (two papers)

- Assume only many bilingual parallel corpora.
- For *N* languages: learn *N* encoders and *N* decoders.
- But what about attention?

$$p(f_i|f_{i-1}, ..., f_1, \mathbf{e}) = g(f_{i-1}, s_i, c_i)$$

$$c_i = \sum_{j=1}^{|\mathbf{e}|} \alpha_{ij} h_j \qquad \alpha_{ij} = \frac{\exp(a_{ij})}{\sum_{k=1}^{|\mathbf{e}|} \exp(a_{ik})}$$

$$a_{ij} = a(s_{i-1}, h_j)$$

Multi-way MT

Firat et al. 2016 (two papers)

- Assume only many bilingual parallel corpora.
- For *N* languages: learn *N* encoders and *N* decoders.
- But what about attention?

$$p(f_i|f_{i-1},...,f_1,\mathbf{e}) = g(\boxed{f_{i-1}, s_i, c_i})$$

Everything we need is right here!

$$\alpha_{ij} = \frac{\exp(a_{ij})}{\sum_{k=1}^{|\mathbf{e}|} \exp(a_{ik})}$$

$$c_i = \sum_{j=1}^{|\mathbf{e}|} \alpha_{ij} h_j$$

$$a_{ij} = a(s_{i-1}, h_j)$$

Multi-way MT

Firat et al. 2016 (two papers)

- As in Bahdanu et al. (2014), attention mechanism is a feedforward function of both decoder hidden state and encoder context vector.

- **Shared** between all encoders and decoders.

$$p(f_i | f_{i-1}, ..., f_1, \mathbf{e}) = g(\boxed{f_{i-1}, s_i, c_i})$$

Everything we need is right here!

$$c_i = \sum_{j=1}^{|\mathbf{e}|} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(a_{ij})}{\sum_{k=1}^{|\mathbf{e}|} \exp(a_{ik})}$$

$$a_{ij} = a(s_{i-1}, h_j)$$

Multi-way MT

Firat et al. 2016 (two papers)

	Size	Single	Single+DF	Multi
En→Fi	100k	5.06/3.96	4.98/3.99	6.2/**5.17**
	200k	7.1/6.16	7.21/6.17	8.84/**7.53**
	400k	9.11/7.85	9.31/8.18	11.09/**9.98**
	800k	11.08/9.96	11.59/10.15	12.73/**11.28**
De→En	210k	14.27/13.2	14.65/13.88	16.96/**16.26**
	420k	18.32/17.32	18.51/17.62	19.81/**19.63**
	840k	21/19.93	21.69/20.75	22.17/**21.93**
	1.68m	23.38/23.01	23.33/22.86	23.86/**23.52**
En→De	210k	11.44/11.57	11.71/11.16	12.63/**12.68**
	420k	14.28/14.25	14.88/15.05	15.01/**15.67**
	840k	17.09/17.44	17.21/17.88	17.33/**18.14**
	1.68m	19.09/19.6	19.36/20.13	19.23/**20.59**

Table 2: BLEU scores where the target pair's parallel corpus is constrained to be 5%, 10%, 20% and 40% of the original size. We report the BLEU scores on the development and test sets (separated by /) by the single-pair model (Single), the single-pair model with monolingual corpus (Single+DF) and the proposed multi-way, multilingual model (Multi).

Low-resource **simulation** (using high-resource European languages)

Multi-way MT

Firat et al. 2016 (two papers)

		Dir	Fr (39m) → En	Fr (39m) En →	Cs (12m) → En	Cs (12m) En →	De (4.2m) → En	De (4.2m) En →	Ru (2.3m) → En	Ru (2.3m) En →	Fi (2m) → En	Fi (2m) En →
(a) BLEU	Dev	Single	27.22	26.91	21.24	15.9	24.13	20.49	21.04	18.06	13.15	9.59
		Multi	26.09	25.04	21.23	14.42	23.66	19.17	21.48	17.89	12.97	8.92
	Test	Single	27.94	**29.7**	20.32	**13.84**	24	**21.75**	22.44	**19.54**	12.24	**9.23**
		Multi	**28.06**	27.88	**20.57**	13.29	**24.20**	20.59	**23.44**	19.39	**12.61**	8.98
(b) LL	Dev	Single	-50.53	-53.38	-60.69	-69.56	-54.76	-61.21	-60.19	-65.81	-88.44	-91.75
		Multi	-50.6	-56.55	-54.46	-70.76	-54.14	-62.34	-54.09	-63.75	-74.84	-88.02
	Test	Single	-43.34	**-45.07**	-60.03	**-64.34**	-57.81	**-59.55**	-60.65	-60.29	-88.66	-94.23
		Multi	**-42.22**	-46.29	**-54.66**	-64.80	**-53.85**	-60.23	**-54.49**	**-58.63**	**-71.26**	**-88.09**

Table 3: (a) BLEU scores and (b) average log-probabilities for all the five languages from WMT'15.

Multi-way MT

Firat et al. 2016 (two papers)

		Dir	Fr (39m) → En	Fr (39m) En →	Cs (12m) → En	Cs (12m) En →	De (4.2m) → En	De (4.2m) En →	Ru (2.3m) → En	Ru (2.3m) En →	Fi (2m) → En	Fi (2m) En →
(a) BLEU	Dev	Single	27.22	26.91	21.24	15.9	24.13	20.49	21.04	18.06	13.15	9.59
		Multi	26.09	25.04	21.23	14.42	23.66	19.17	21.48	17.89	12.97	8.92
	Test	Single	27.94	**29.7**	20.32	**13.84**	24	**21.75**	22.44	**19.54**	12.24	**9.23**
		Multi	**28.06**	27.88	**20.57**	13.29	**24.20**	20.59	**23.44**	19.39	**12.61**	8.98
(b) LL	Dev	Single	-50.53	-53.38	-60.69	-69.56	-54.76	-61.21	-60.19	-65.81	-88.44	-91.75
		Multi	-50.6	-56.55	-54.46	-70.76	-54.14	-62.34	-54.09	-63.75	-74.84	-88.02
	Test	Single	-43.34	**-45.07**	-60.03	**-64.34**	-57.81	**-59.55**	-60.65	-60.29	-88.66	-94.23
		Multi	**-42.22**	-46.29	**-54.66**	-64.80	**-53.85**	-60.23	**-54.49**	**-58.63**	**-71.26**	**-88.09**

Table 3: (a) BLEU scores and (b) average log-probabilities for all the five languages from WMT'15.

ok, but what about multi-source?

Multi-way multi-source MT

Firat et al. 2016 (two papers)

- Still assumes only many bilingual parallel corpora.
- What to do if there are multiple input sentences?
- Early averaging (average context vectors). $\mathbf{c}_t = \frac{\mathbf{c}_t^1 + \mathbf{c}_t^2}{2}.$
- Late averaging (aka linear interpolation).

$$P(w_i|\boldsymbol{c}) = \sum_{k=1}^{K} \lambda_k(\boldsymbol{c}) P_k(w_i|\boldsymbol{c})$$

Early and late averaging are orthogonal, can be combined.

Multi-way multi-source MT

Firat et al. 2016 (two papers)

	Src	Trgt	Multi Test	Single Test
(a)	Es	En	28.32	27.48
(b)	Fr	En	27.93	27.21
(c)	En	Es	28.41	28.90
(d)	En	Fr	23.41	24.05

Table 2: One-to-one translation qualities using the multi-way, multilingual model and four separate single-pair models.

Multi-way multi-source MT

Firat et al. 2016 (two papers)

	Src	Trgt	Multi Test	Single Test
(a)	Es	En	28.32	27.48
(b)	Fr	En	27.93	27.21

Table 2: One-to-one translation qualities using the multi-way multilingual model and four separate single-pair models.

		Multi		Single	
		Dev	Test	Dev	Test
(a)	Early	31.89	31.35	–	–
(b)	Late	32.04	31.57	32.00	31.46
(c)	E+L	32.61	31.88	–	–

Table 3: Many-to-one quality (Es+Fr→En) using three translation strategies. Compared to Table 2 (a–b) we observe a significant improvement (up to 3+ BLEU), although the model was never trained in these many-to-one settings. The second column shows the quality by the ensemble of two separate single-pair models.

Multi-way multi-source MT

Firat et al. 2016 (two papers)

	Src	Trgt	Multi Test	Single Test
(a)	Es	En	28.32	27.48
(b)	Fr	En	27.93	27.21

Table 2: One-to-one translation qualities using the multi-way multilingual model and four separate single-pair models.

		Multi Dev	Multi Test	Single Dev	Single Test
(a)	Early	31.89	31.35	–	–
(b)	Late	32.04	31.57	32.00	31.46
(c)	E+L	32.61	31.88	–	–

Table 3: Many-to-one quality (Es+Fr→En) using three translation strategies. Compared to Table 2 (a–b) we observe a significant improvement (up to 3+ BLEU), although the model was never trained in these many-to-one settings. The second column shows the quality by the ensemble of two separate single-pair models.

Zero-shot MT

Firat et al. 2016 (two papers)

- Suppose our bilingual parallel data include a pair of languages for which we have no parallel data.

Spanish $\longleftrightarrow$ English   English $\longleftrightarrow$ French

- Q: Can we use the multi-way encoder-decoder system to translate Spanish into French?

Zero-shot MT

Firat et al. 2016 (two papers)

- Suppose our bilingual parallel data include a pair of languages for which we have no parallel data.
- Q: Can we use the multi-way encoder-decoder system to translate Spanish into French?

	Pivot	Many-to-1	Dev	Test
(a)			< 1	< 1
(b)	$\surd$		20.64	20.4

Table 4: Zero-resource translation from Spanish (Es) to French (Fr) *without* finetuning. When pivot is $\surd$, English is used as a pivot language.

A: Not really

Must pivot (explicitly) through English

Zero-shot MT

Firat et al. 2016 (two papers)

- *Finetuning*: what if we use a small amount of parallel data in this setting?
- Q: Where would we get this data?

	Pivot	Many-to-1	Dev	Test
(a)			< 1	< 1
(b)	√		20.64	20.4

Table 4: Zero-resource translation from Spanish (Es) to French (Fr) *without* finetuning. When pivot is √, English is used as a pivot language.

Zero-shot MT

Firat et al. 2016 (two papers)

- *Finetuning*: what if we use a small amount of parallel data in this setting?
- Q: Where would we get this data? **Backtranslation**

Spanish $\longleftrightarrow$ English English $\longleftrightarrow$ French

Zero-shot MT

Firat et al. 2016 (two papers)

- *Finetuning*: what if we use a small amount of parallel data in this setting?
- Q: Where would we get this data? **Backtranslation**

$$\text{Spanish} \longleftrightarrow \text{English} \qquad \begin{array}{ccc} \text{English} & \longleftrightarrow & \text{French} \\ \updownarrow & & \\ \text{Spanish (MT)} & & \end{array}$$

Zero-shot MT

Firat et al. 2016 (two papers)

- *Finetuning*: what if we use a small amount of parallel data in this setting?

- Q: Where would we get this data? **Backtranslation**

Spanish ⟷ English

English ⟷ French

Spanish (MT)

Zero-shot MT

Firat et al. 2016 (two papers)

- *Finetuning*: what if we use a small amount of parallel data in this setting?

Pivot	Many-to-1		Pseudo Parallel Corpus 1k	10k	100k	1m
Single-Pair Models		Dev	–	–	–	–
		Test	–	–	–	–
√	No Finetuning		Dev: 20.64, Test: 20.4			
		Dev	0.28	10.16	15.61	17.59
		Test	0.47	10.14	15.41	17.61

Zero-shot MT

Firat et al. 2016 (two papers)

- *Finetuning*: what if we use a small amount of parallel data in this setting?

			Pseudo Parallel Corpus				True Parallel Corpus			
Pivot	Many-to-1		1k	10k	100k	1m	1k	10k	100k	1m
Single-Pair Models		Dev	–	–	–	–	–	–	11.25	21.32
		Test	–	–	–	–	–	–	10.43	20.35
√	No Finetuning		Dev: 20.64, Test: 20.4				–			
		Dev	0.28	10.16	15.61	17.59	0.1	8.45	16.2	20.59
		Test	0.47	10.14	15.41	17.61	0.12	8.18	15.8	19.97

Zero-shot MT

Johnson et al. 2016 (Google)

- Incremental training: add a small amount of (true) parallel data in the language pair of interest.

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

	Model	BLEU
(a)	PBMT bridged	28.99
(b)	NMT bridged	30.91
(c)	NMT Pt→Es	31.50
(d)	Model 1 (Pt→En, En→Es)	21.62
(e)	Model 2 (En↔{Es, Pt})	24.75
(f)	Model 2 + incremental training	31.77

Zero-shot MT

Johnson et al. 2016 (Google)

Table 6: BLEU scores for English↔{Belarusian, Russian, Ukrainian} models.

	Zero-Shot	From-Scratch	Incremental
English→Belarusian	16.85	17.03	16.99
English→Russian	22.21	22.03	21.92
English→Ukrainian	18.16	17.75	18.27
Belarusian→English	25.44	24.72	25.54
Russian→English	28.36	27.90	28.46
Ukrainian→English	28.60	28.51	28.58
Belarusian→Russian	56.53	82.50	78.63
Russian→Belarusian	58.75	72.06	70.01
Russian→Ukrainian	21.92	25.75	25.34
Ukrainian→Russian	16.73	30.53	29.92

trained on parallel data

Zero-shot MT

Johnson et al. 2016 (Google)

Table 6: BLEU scores for English↔{Belarusian, Russian, Ukrainian} models.

	Zero-Shot	From-Scratch	Incremental
English→Belarusian	16.85	17.03	16.99
English→Russian	22.21	22.03	21.92
English→Ukrainian	18.16	17.75	18.27
Belarusian→English	25.44	24.72	25.54
Russian→English	28.36	27.90	28.46
Ukrainian→English	28.60	28.51	28.58
Belarusian→Russian	56.53	82.50	78.63
Russian→Belarusian	58.75	72.06	70.01
Russian→Ukrainian	21.92	25.75	25.34
Ukrainian→Russian	16.73	30.53	29.92

actual zero-shot experiment

Zero-shot MT

Johnson et al. 2016 (Google)

code-switching in the input language:

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.

code-switching in the output language:

Spanish/Portuguese:	Here the other guinea-pig cheered, and was suppressed.
$w_{pt} = 0.00$	Aquí el otro conejillo de indias animó, y fue suprimido.
$w_{pt} = 0.30$	Aquí el otro conejillo de indias animó, y fue suprimido.
$w_{pt} = 0.40$	Aquí, o outro porquinho-da-índia alegrou, e foi suprimido.
$w_{pt} = 0.42$	Aqui o outro porquinho-da-índia alegrou, e foi suprimido.
$w_{pt} = 0.70$	Aqui o outro porquinho-da-índia alegrou, e foi suprimido.
$w_{pt} = 0.80$	Aqui a outra cobaia animou, e foi suprimida.
$w_{pt} = 1.00$	Aqui a outra cobaia animou, e foi suprimida.

Zero-shot MT

Johnson et al. 2016 (Google)

Portuguese informant: “we decided it's impossible to judge the correctness of the translation without context (but it's likely wrong). After finding the context (Alice in Wonderland) we can conclude it's wrong.”

code-switching in the output language:

Spanish/Portuguese:	Here the other guinea-pig cheered, and was suppressed.
$w_{pt} = 0.00$	Aquí el otro conejillo de indias animó, y fue suprimido.
$w_{pt} = 0.30$	Aquí el otro conejillo de indias animó, y fue suprimido.
$w_{pt} = 0.40$	Aquí, o outro porquinho-da-índia alegrou, e foi suprimido.
$w_{pt} = 0.42$	Aqui o outro porquinho-da-índia alegrou, e foi suprimido.
$w_{pt} = 0.70$	Aqui o outro porquinho-da-índia alegrou, e foi suprimido.
$w_{pt} = 0.80$	Aqui a outra cobaia animou, e foi suprimida.
$w_{pt} = 1.00$	Aqui a outra cobaia animou, e foi suprimida.

Google Interlingua (Johnson et al., 2016)

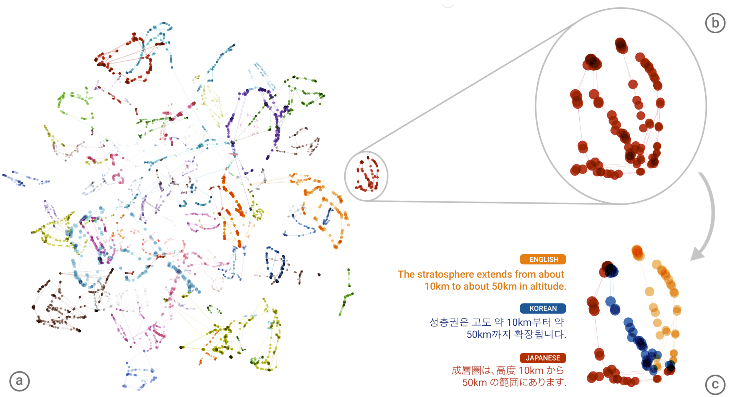


Figure 2: A t-SNE projection of the embedding of 74 semantically identical sentences translated across all 6 possible directions, yielding a total of 9,978 steps (dots in the image), from the model trained on English↔Japanese and English↔Korean examples. (a) A bird's-eye view of the embedding, coloring by the index of the semantic sentence. Well-def ned clusters each having a single color are apparent. (b) A zoomed in view of one of the clusters with the same coloring. All of the sentences within this cluster are translations of "The stratosphere extends from about 10km to about 50km in altitude." (c) The same cluster colored by source language. All three source languages can be seen within this cluster.

Interlingua?

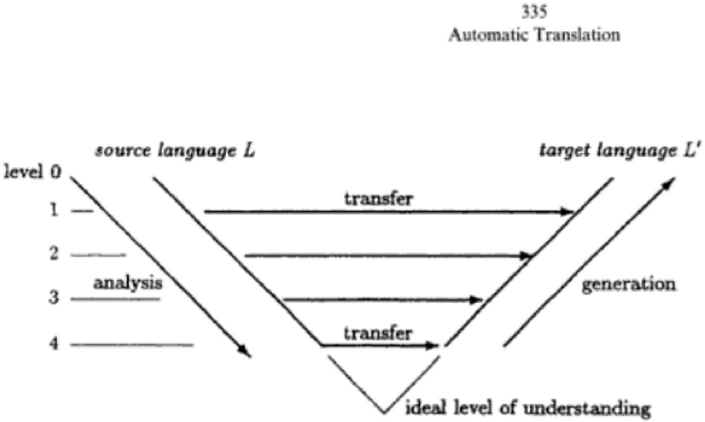


Figure 28.1

From Vauquois (1968), reproduced by Adam Lopez.

Interlingua?

- Theoretically, a very inspiring concept.
- Need for $2N$ instead of n^2 systems.
- Sceptical view:
  - Need to capture all distinctions in word meanings: `https://en.wikipedia.org/wiki/Eskimo_words_for_snow`
  - Text form underspecifies the meaning, formally captured content underspecifies the form (Lampert, 2001).
  - Interannotator agreement decreases as we proceed along layers of linguistic analysis (Dorr et al., 2010).

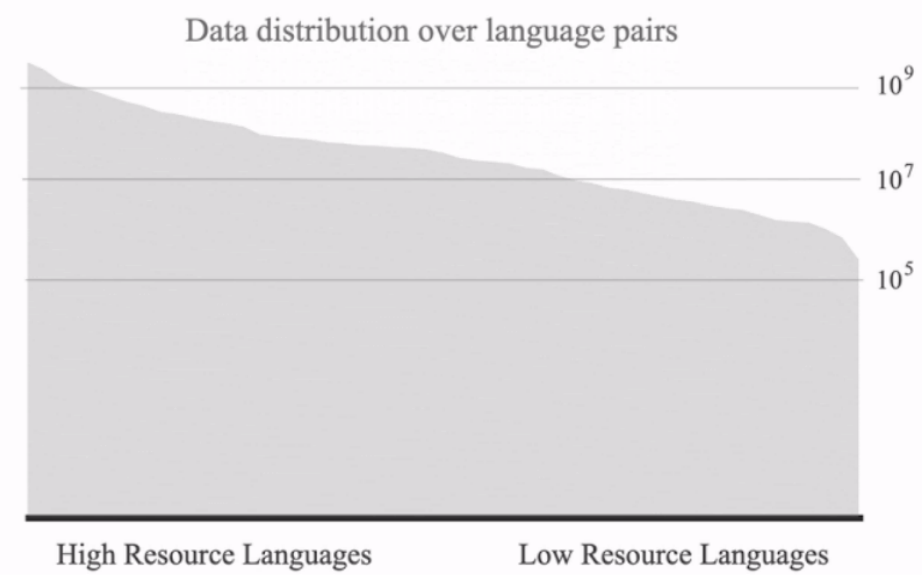
Interlingua?

- Optimistic/wishful view:
  - Molto-Project (EU FP7, 2011-2013), among others: `http://www.molto-project.eu/`

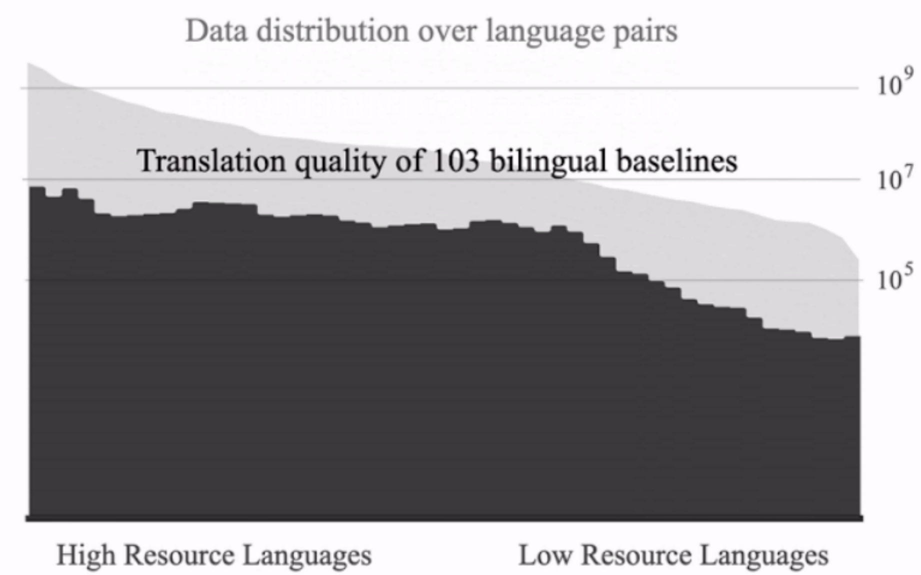
Isn't interlingua an unrealistic dream? Yes, it is, if we want to have a universal interlingua working for everything. This is why we don't believe we can ever translate newspapers with MOLTO techniques. However, domain-specific interlinguas have proved quite feasible. Notice that this move is similar to what has happened in ontologies: they have moved from universal ontologies to domain ontologies.

Massively Multi-Lingual Models

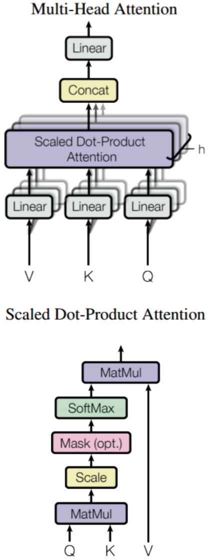
Available Data for EN↔100+ Langs

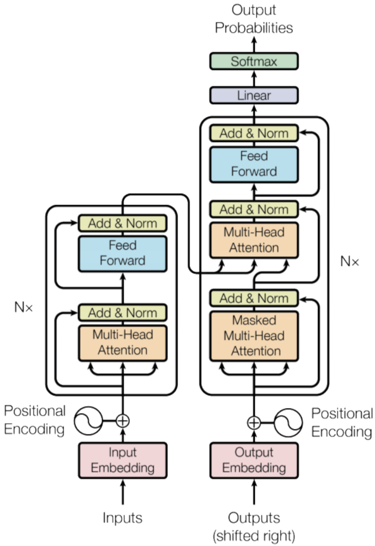


Translation Quality of Bilingual MT



Standard Transformer Model





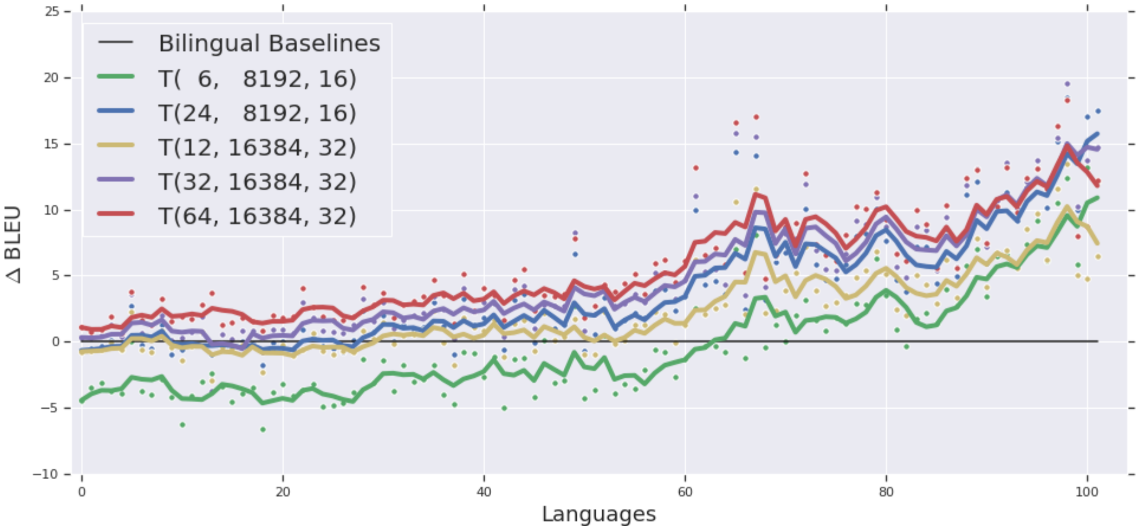
Google Transformer Sizes

GPipe (Huang et al., 2019) introduces microbatches for faster training of deep models across multiple GPUs.

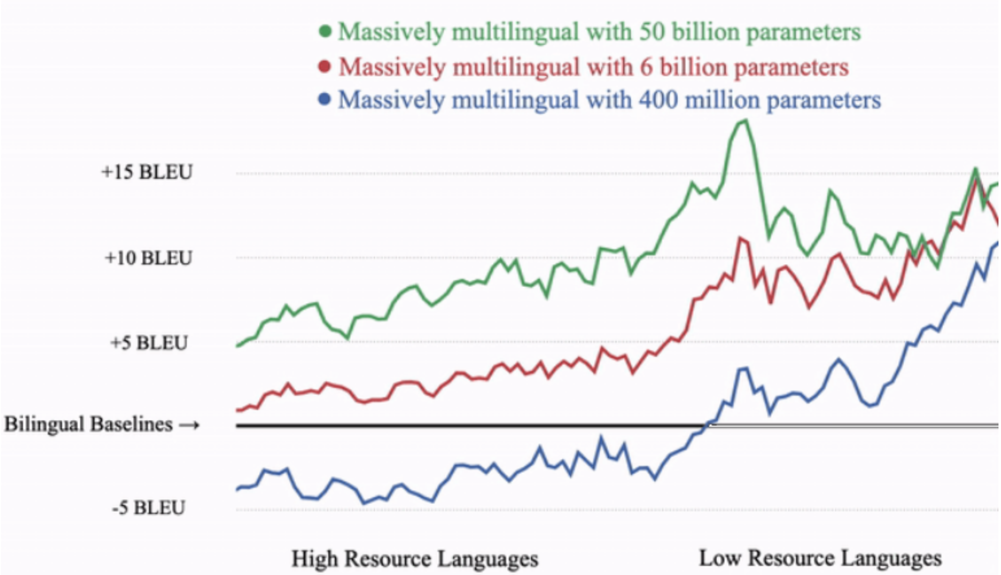
Enc/Dec Depth	FF Dim	Heads	Total Parameters	GPUs Used	
6	8192	16	400M	1 default	
12	16384	32	1.3B	2	"wide"
24	8192	16	1.3B	4	"deep"
32	16384	32	3.0B	8	
64	16384	32	6.0B	16	

- "Deep" better than "wide" on low-resource languages.
    - Indicates better generalization.
- Further tricks needed to keep the training stable.

Massively Multilingual Models



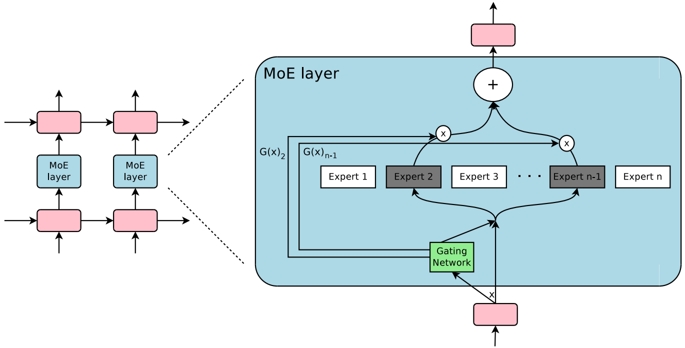
Massive Massively Multilingual Models



Google-Sized Experiment

The recent 50 billion parameters Transformer needed further trick:

- sparsely-gated mixture of experts (Shazeer et al., 2017):

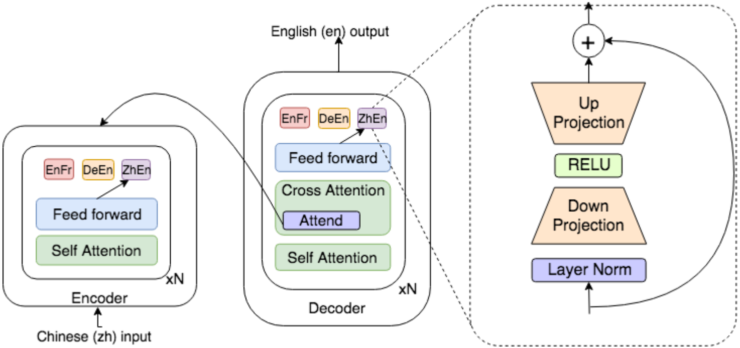


⇒ BLEU on 100 langs re-gained and improved by 125x larger model.

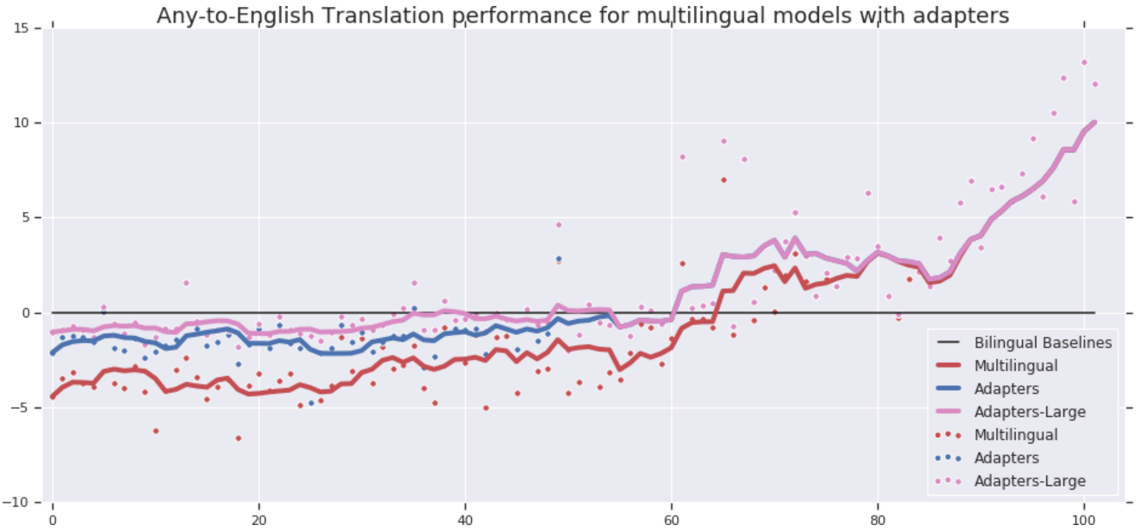
https://ai.googleblog.com/2019/10/exploring-massively-multilingual.html

Domain Adapters to Recover Practical Sizes

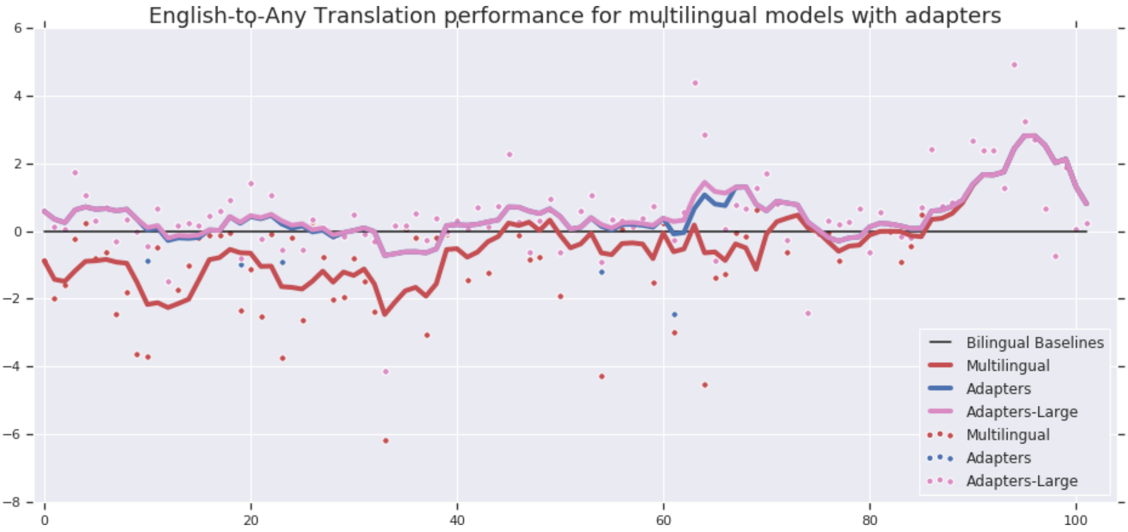
- Bapna and Firat (2019) propose tiny tunable "adapter" layers.
  1. Pretrain on a large mixed-language corpus.
  2. Inject adapter layers.
  3. Finetune adapter layers for each of the target tasks.



Domain Adapters into English



Domain Adapters from English



Summary

- Tools from machine translation for reuse in multilingual research.
- Machine translation is multilingual from the beginning.
- Transfer learning in NMT works.
  - $\Rightarrow$ NMT can exploit more and less related data.
  - Trivial Transfer: Parent just has to be larger.
  - Even unrelated language pairs can help.
  - Very big improvements in low-resource conditions.
- Language families emerge in language token embedding.
- Model capacity is the bottleneck.
  - Models 125x large for 100 languages in one model allow gains on high-resource languages, too.
  - With tiny adaptors instead of mixture of experts model sizes can decrease again.

References

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1538–1548, Hong Kong, China, November. Association for Computational Linguistics.

Fabienne Braune and Alexander Fraser. 2010. Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora. In Coling 2010: Posters, pages 81–89, Beijing, China, August. Coling 2010 Organizing Committee.

Bonnie j. Dorr, Rebecca j. Passonneau, David Farwell, Rebecca Green, Nizar Habash, Stephen Helmreich, Eduard Hovy, Lori Levin, Keith j. Miller, Teruko Mitamura, Owen Rambow, and Advaith Siddharthan. 2010. Interlingual annotation of parallel text corpora: A new framework for annotation and evaluation. Nat. Lang. Eng., 16(3):197–243.

Miquel Esplà-Gomis and Mikel L. Forcada. 2010. Combining Content-Based and URL-Based Heuristics to Harvest Aligned Bitexts from Multilingual Sites with Bitextor. In Prague Bulletin of Mathematical Linguistics - Special Issue on Open Source Machine Translation Tools, number 93 in Prague Bulletin of Mathematical Linguistics. Charles University, January.

William A. Gale and Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics, 19(1):75–102.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2017. Effective strategies in zero-shot neural machine translation. CoRR, abs/1711.07893.

Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, and zhifeng Chen. 2019. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 103–112. Curran Associates, Inc.