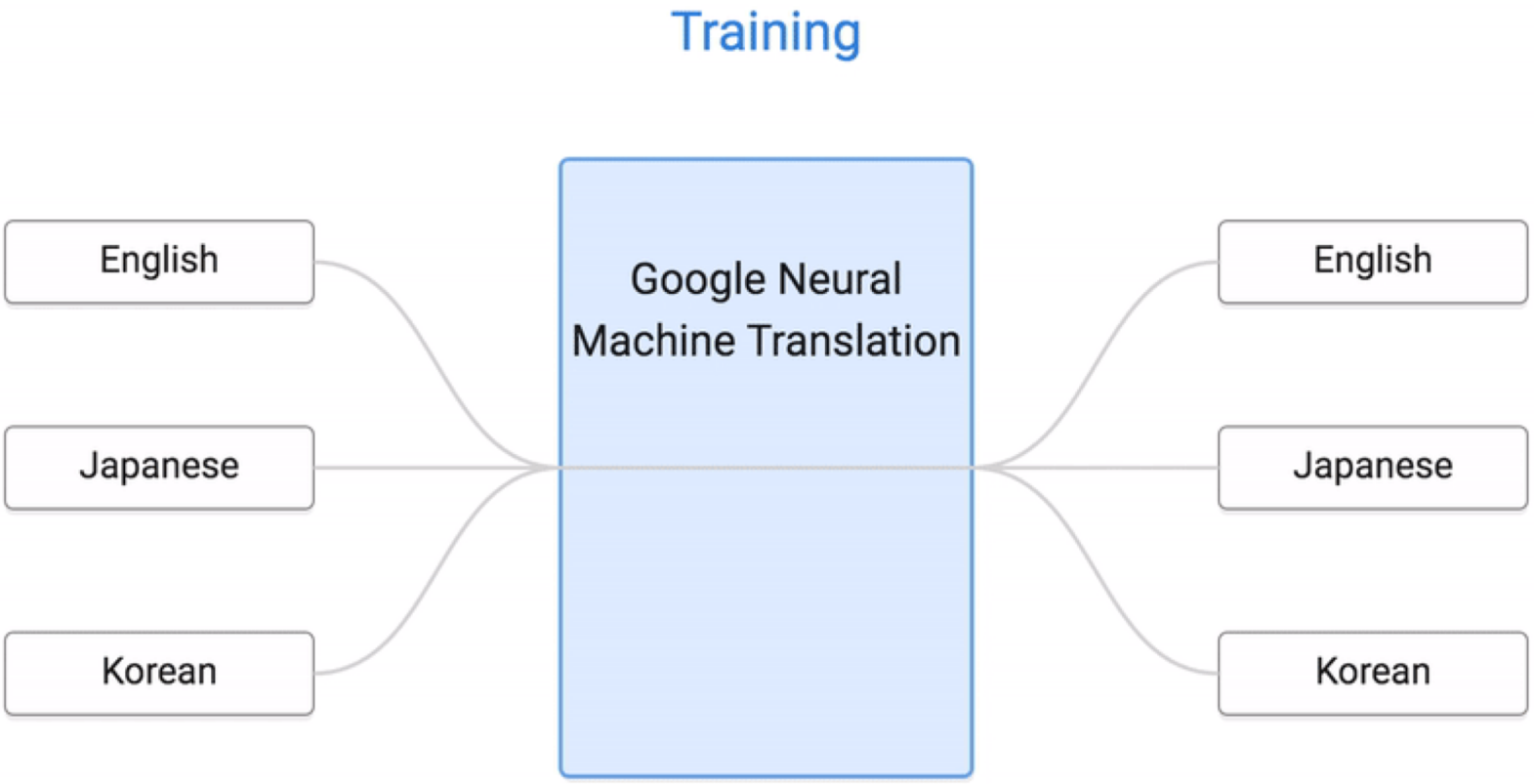
Translational grounding: Multilingual NMT for paraphrasing

Yves Scherrer & Jörg Tiedemann

University of Helsinki

Zero-shot machine translation



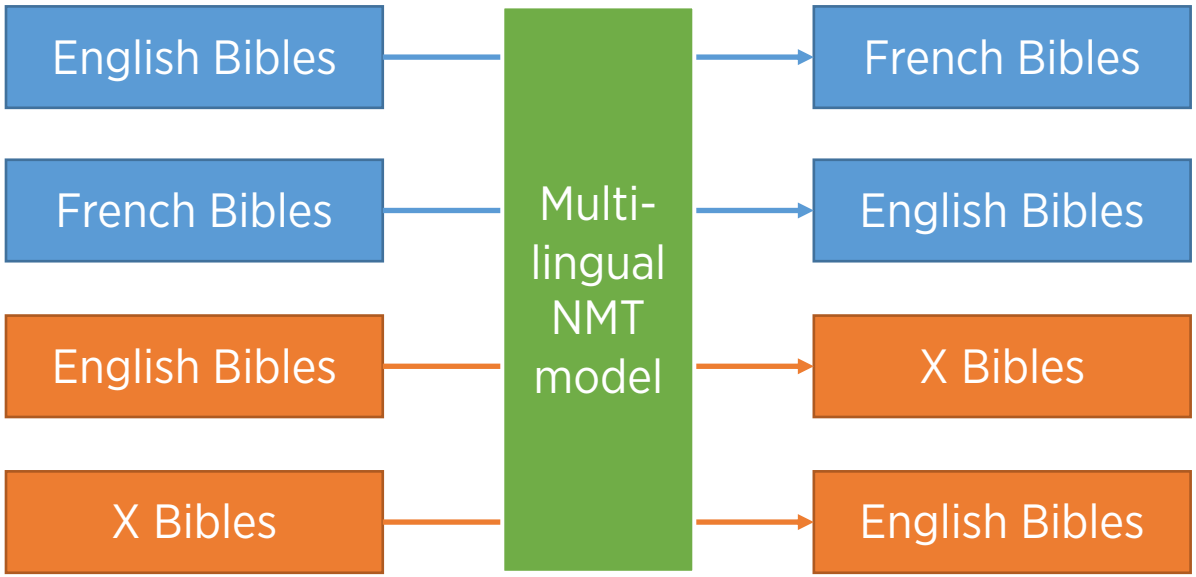
https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html

Johnson et al.: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. TACL 5, 2017.

What if we tell the model to translate from English to English?

- Hypotheses:
  - The model will generate paraphrases of the original sentences
  - Adding more languages will lead to better abstraction capabilities of the model...
  - ... and therefore, to better paraphrase generation and recognition

Experimental setup



BASE model

+X models

X ∈ Afrikaans, Albanian, Breton, German, Greek, Frisian, Hindi, Italian, Dutch, Ossetian, Polish, Russian, Slovene, Spanish, Serbian, Swedish, ALL

Two test sets:

- Held-out English **Bible** data (in-domain)
- **Tatoeba** crowdsourced paraphrases (out-of-domain)

Paraphrase generation

Source	He slept soundly.
BASE	Et il se prosterna devant soi.
+Breton	And, behold, he rose up quickly.
+German	And he began to sleep.
+Greek	He was sleeping.
+ALL	And when he had died, he was asleep.

Although we told the model to generate English, it insisted on generating French…

Source	She has no brothers.
BASE	Elle n'a point de frères.
+Breton	Or, elle n'a pas de frères.
+German	For she has no brothers.
+Ossetian	No, brothers.
+ALL	You have no brothers.

Evaluation of generated paraphrases

- BLEU with source = reference
  - High values = large overlap
  - Maximum values obtained trivially by copying
- PINC (Chen & Dolan 2011)
  - High values = large diversity
  - Maximum values obtained trivially by generating random text
- Good paraphrases should have high BLEU as well as high PINC scores

Evaluation of generated paraphrases

- Will the model not just learn to copy the input?

- Percentages of identical source and target sentences:

Model	Bible	Tatoeba
English–French	0.0%	0.7%
+ Afrikaans	0.9%	4.8%
+ Albanian	0.7%	3.4%
+ Breton	0.0%	1.1%
+ German	1.4%	4.9%
+ Greek	1.1%	5.2%
+ Frisian	0.7%	4.3%
+ Hindi	0.9%	4.2%
+ Italian	1.2%	5.0%
+ Dutch	1.1%	5.1%
+ Ossetian	0.6%	3.5%
+ Polish	0.4%	2.8%
+ Russian	1.4%	4.7%
+ Slovene	0.6%	3.2%
+ Spanish	1.1%	5.5%
+ Serbian	0.5%	3.3%
+ Swedish	1.2%	4.9%
+ All	0.8%	2.0%
+ English–English	71.6%	70.0%

Evaluation of generated paraphrases

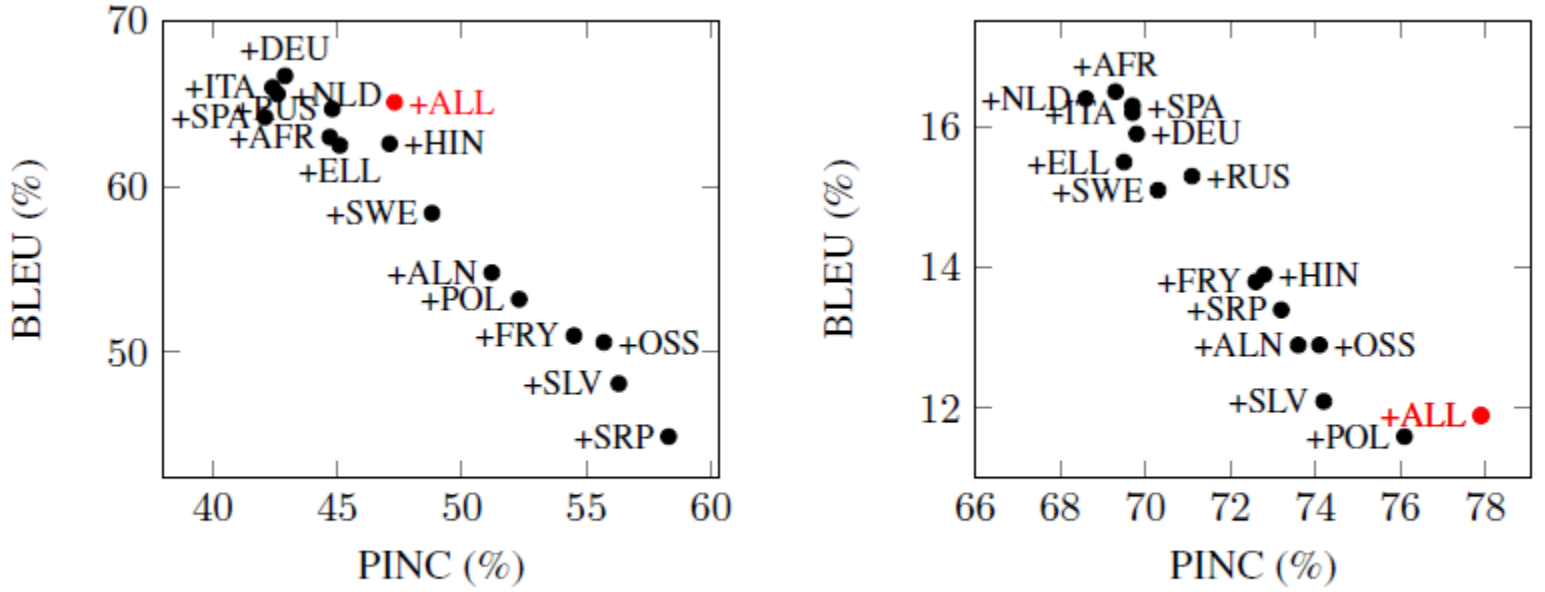
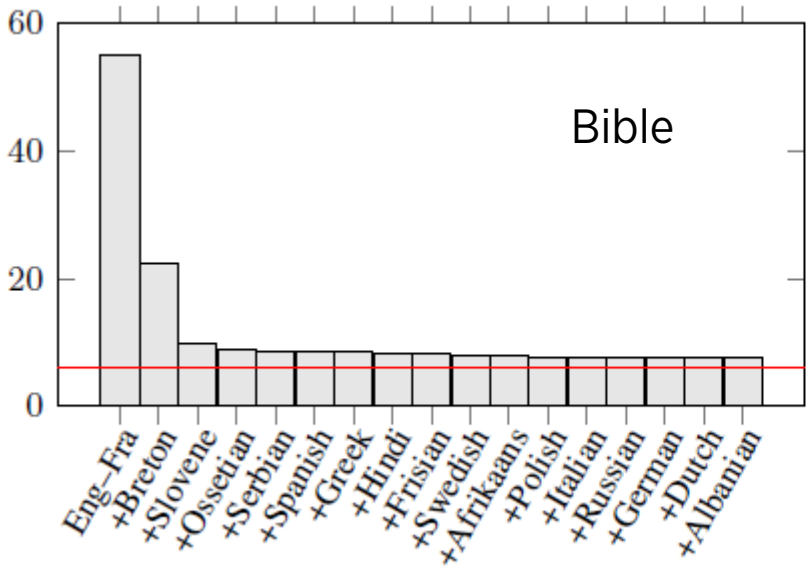
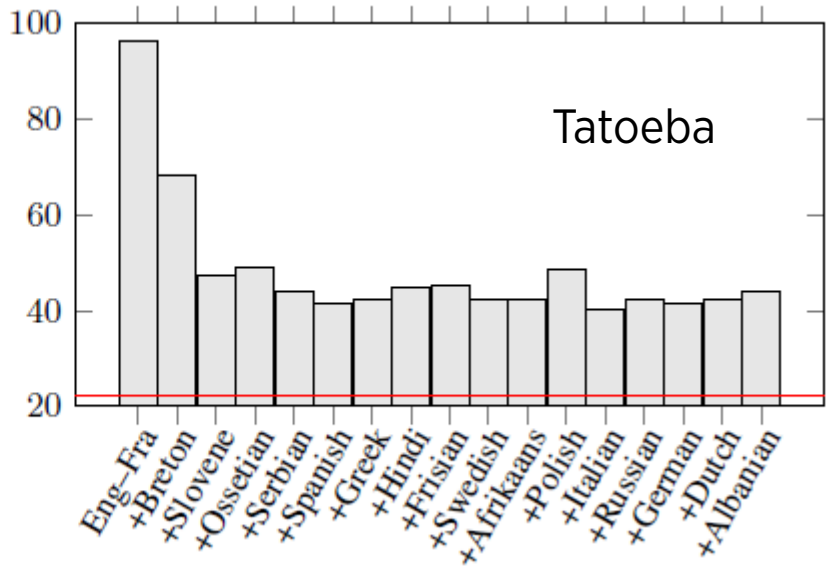


Figure 6: Paraphrase BLEU vs. PINC scores for the Bible test set (left) and the Tatoeba test set (right).

Paraphrase recognition

- Ask the model to score given paraphrases
- Intuition: Models with higher level of semantic abstraction should obtain lower perplexity when seeing paraphrased sentences





Domain effects

- What happens to the contemporary vocabulary of the Tatoeba test set when paraphrased by a Bible-trained model?

Source	Do you have a cellphone?
+Hindi	Do you have a scorpion?

Source	Birds fly.
Base	“Do not go out.
+Albanian	Bear.
+German	The flying creatures shall fly away.
+Greek	Blind guides!
+ALL	They run quickly.

Source	Have you never eaten a kiwi?
+Afrikaans	Have you not eaten sour grapes?

Source	Could I park my car here?
+Italian	Do I get up here with my cavalry?

Source	Do your children speak French?
+Spanish	Do your children speak Greek?

Conclusions

- Multilingual models build more abstract semantic representations
  - Striking difference between bilingual base model and trilingual +X models
  - Quantitative, but no qualitative benefit from additional languages (+ALL model)
  - Size and variety of training data has larger effect than language family
- Future work:
  - Discriminating paraphrases from non-paraphrases
  - More principled analysis of models