Massively Multilingual Neural Machine Translation

Roee Aharoni, Melvin Johnson, Orhan Firat

roeeaharoni,melvinp,orhanf@google.com

Hey! What is this paper about?

Imagine a single universal NMT model, that can translate in more than 100 languages...

Well, does it work?

Yes, quite well actually - in a low resource setting we got great results with a 59-language many-to-many model. See the **green panel** for more.

That's interesting! But does it scale? Larger datasets? More languages?

We then scaled to 103 languages, with one million examples per pair. It still works well, and outperforms bilingual baselines - see the **red panel**.

Did you do any analysis?

Our **ablation** shows that in the high resource case, adding more languages can harm performance - so we may need larger models...

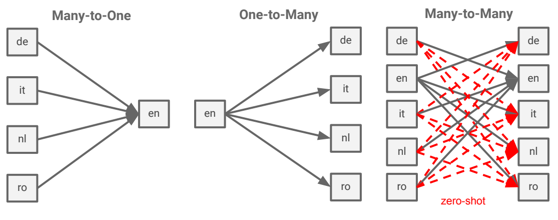
And what about zero-shot performance?

Here we actually saw an opposite trend - adding more languages helps generalization in zero-shot directions.

How can I learn more about this?

Read our paper using the QR code ⟶

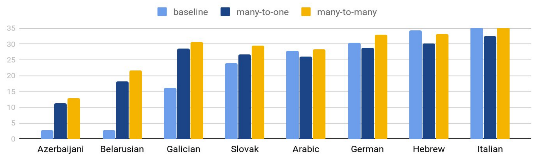
Multilingual Model Types



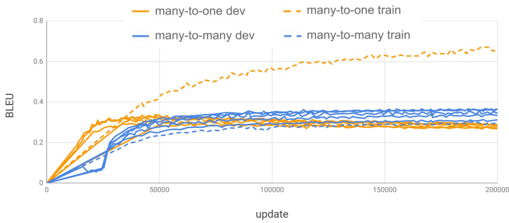
Low Resource NMT

- TED talks corpus (Qi et al. 2018)
- English-centric - 58 languages ↔ English
- Highly imbalanced - 4k to 200k examples per pair

X→English BLEU



- Many-to-many > many-to-one
- Many-to-many > bilingual baselines (similar capacity)
- More than 1.4 BLEU improvement over the best published result (Wang et al. 18)
- Why many-to-many is better despite a larger burden?

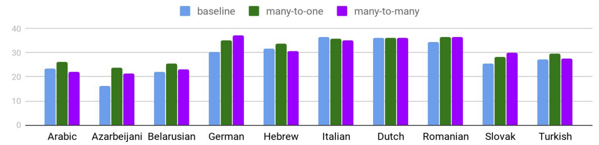


- Many-to-one models suffer from memorization
- Having multiple target languages prevents such memorization and improves performance
- Other direction - English-to-Many models are better as there is no English bias

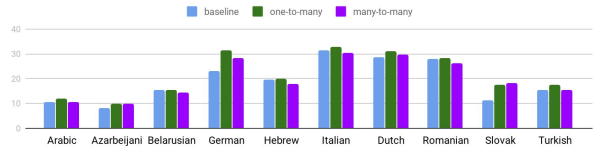
Towards Universal NMT

- 102 languages ↔ English
- 1 million examples per direction
- 204 directions (language-pairs/tasks)
- Large transformer model - 473M parameters

X → English BLEU



English → X BLEU

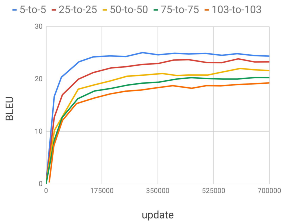


- In the high-resource case, many-to-one and one-to-many models win (capacity bottleneck)
- Training becomes unstable ("interference")

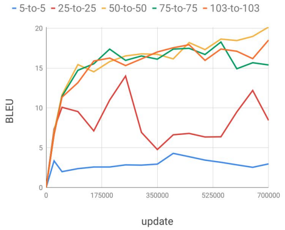
Analysis

- What if we vary the number of languages?

Supervised (En→Ru)



Zero-Shot (Uk→Ru)



- Supervised directions deteriorate with more languages - capacity bottleneck
- Zero-shot improves with more languages - better generalization