Breaking Down Multilingual Machine Translation

Ting-Rui Chiang\textsuperscript{1} Yi-Pei Chen\textsuperscript{2} Yi-Ting Yeh\textsuperscript{1} Graham Neubig\textsuperscript{1}
Carnegie Mellon University\textsuperscript{1}, The University of Tokyo\textsuperscript{2}
Background: Multilingual Training for Machine Translation

We were wondering why there is more improvement in one direction than the other.

encoder -> decoder

many-to-one

↑ has more improvement than ↓
Observation

Azerbaijani
Belarusian
Arabic
German

encoder
decoder

English

many-to-one

affects the number of modalities exposed to the encoder/decoder

English
encoder
decoder

one-to-many

Azerbaijani
Belarusian
Arabic
German
Investigation

- How does multilingual training affect the encoder/decoder?
  - i.e. How useful are the parameters learned from multilingual training?
Experiment - Step 1: Train a Multilingual Model

- Azerbaijani
- Belarusian
- Arabic
- German

encoder

decoder

train

parameters
Experiment - Step 2: Initialize Several Bilingual Models

Encoder | Decoder
---|---
English | Arabic

Encoder | Decoder
---|---
English | Arabic

Encoder | Decoder
---|---
English | Arabic

Encoder | Decoder
---|---
English | Arabic
Experiment - Step 2: Initialize Several Bilingual Models

Load both encoder and decoder parameters.
Experiment - Step 2: Initialize Several Bilingual Models

- **Load both encoder and decoder parameters**
  - English to Arabic
- **Load encoder**
  - English to Arabic
  - English to Arabic
  - English to Arabic
Experiment - Step 2: Initialize Several Bilingual Models

Load both
- English
- Arabic

Load encoder
- English
- Arabic

Load decoder
- English
- Arabic
Experiment - Step 2: Initialize Several Bilingual Models

- **Load both**
  - English Encoder → Arabic Decoder

- **Load encoder**
  - English Encoder → Arabic Decoder

- **Load decoder**
  - English Encoder → Arabic Decoder

- **From scratch**
  - English Encoder → Arabic Decoder
Experiment - Step 3: Train with Bilingual Data

- **Load both**
  - English: Encoder → Decoder (Arabic)

- **Load encoder**
  - English: Encoder → Decoder (Arabic)

- **Load decoder**
  - English: Encoder → Decoder (Arabic)

- **From scratch**
  - English: Encoder → Decoder (Arabic)
Experiment - Final Step: Compare their performance

We can infer how multilingual training benefits the encoder/decoder.
Low-resource: Multilingual training benefits both the encoder and the decoder.
High-resource: Multilingual training only benefits encoder.
Investigating Parameter Sharing

1. Identify important attention heads for languages.
2. Compute the coherence of important heads.
Investigating Parameter Sharing
## Improvement by Training with Related Languages

| Model                                           | az  | be  | gl  | sk  | ar  | de  | he  | it  |
|------------------------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|
| En-All (Aharoni et al., 2019)                   | 5.1 | 10.7| 26.6| 24.5| 16.7| 30.5| 27.6| 35.9|
| Bilingual Baseline                              | 1.3 | 1.9 | 3.9 | 13.1| 15.6| 27.1| 25.4| 32.0|
| All-All                                         | 3.1 | 6.2 | 20.5| 18.4| 12.7| 24.5| 21.1| 30.5|
| All-All w/ f.t. on related clusters             | 7.9 | 12.8| 27.5| 24.9| -   | 30.2| 27.0| 35.4|
| All-All w/ f.t. on random groups                | 6.9 | 13.3| 22.5| 24.3| -   | -   | 27.5| 35.2|
| En-All                                          | 4.9 | 9.0 | 24.2| 21.9| 15.1| 27.9| 24.1| 33.3|
| En-All w/ f.t. on related clusters              | **7.9** | 13.9| 21.0| **26.2** | 16.7| 30.4| 27.1| 35.4|
| En-All w/ f.t. on random groups                 | 7.0 | 13.1| 23.1| 24.7| -   | -   | 27.6| 35.2|
| Load En-All w/ f.t. on closest                  | 7.8 | **15.2** | **28.6** |
Conclusion

We found that multilingual training is more useful for the encoder.

We proposed a purely data-driven way to identify related languages.

Our experiments can serve as analysis tools for future research.