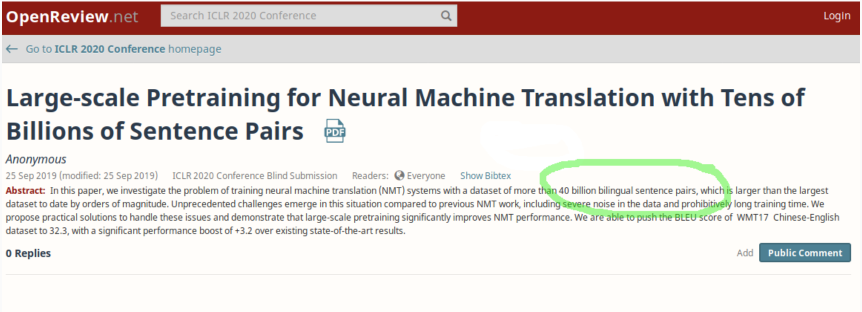
Global Under-Resourced Media Translation

MetaForum

Barry Haddow

October 8th, 2019

Sometimes You Have Lots of Data ...


OpenReview.net
Search ICLR 2020 Conference
Login
← Go to ICLR 2020 Conference homepage

Large-scale Pretraining for Neural Machine Translation with Tens of Billions of Sentence Pairs

Anonymous

25 Sep 2019 (modified: 25 Sep 2019) ICLR 2020 Conference Blind Submission Readers: Everyone Show Bibtex

Abstract: In this paper, we investigate the problem of training neural machine translation (NMT) systems with a dataset of more than 40 billion bilingual sentence pairs, which is larger than the largest dataset to date by orders of magnitude. Unprecedented challenges emerge in this situation compared to previous NMT work, including severe noise in the data and prohibitively long training time. We propose practical solutions to handle these issues and demonstrate that large-scale pretraining significantly improves NMT performance. We are able to push the BLEU score of WMT17 Chinese-English dataset to 32.3, with a significant performance boost of +3.2 over existing state-of-the-art results.

0 Replies

But what if you don't have 40 billion parallel sentences?

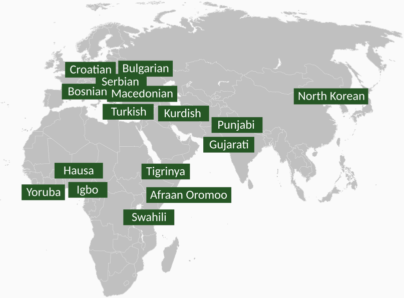
Global Under-Resourced Media Translation

Rationale

- MT is still poor for most world languages

Aims

- Improve translation quality
- Apply to journalism and media analyst use-cases



https://gourmet-project.eu/

Project Overview

Techniques

- Data Gathering and Augmentation
- Modelling Morphological Structure
- Structure Induction at the Sentence Level
- Transfer Learning

Research Partners

Universitat d'Alacant
Universidad de Alicante

User Partners

Data Gathering and Augmentation

Aims

- Collection of corpora, lexical and linguistic resources (including web-scraping for low-resource)
- Data augmentation to extend corpora, applying rule-based techniques

Progress

- Delivered report on resources for all proposed languages
- Sponsored English-Gujarati task at WMT19
  - → Released new parallel and monolingual training corpoa
  - → Also development and test sets
- Sponsoring English-Tamil task at WMT20

Modelling Challenges

Challenges

- Lack of data
- Structurally distant languages
- Morphological complexity and agglutination

Approaches

- Unsupervised and semi-supervised approaches
- Using linguistic and lexical resources
- Joint modelling of alignment and morphology
- Modelling of latent structure
- Better explanation of data via joint source-target modelling

Transfer Learning: Low-resource Systems at WMT19

English↔Gujarati

- Exploit large hi-en corpus
- Build unsupervised hi-gu system
- Using transliteration

  … and similarity

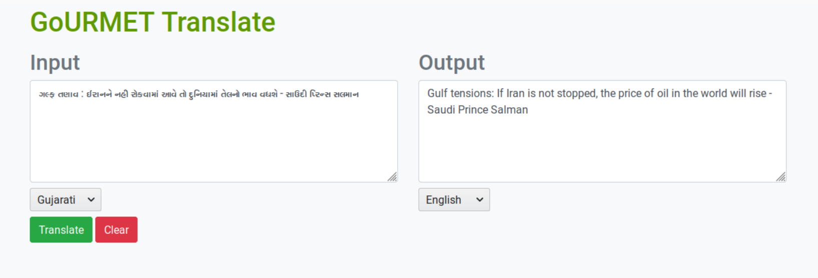
→ Synthesise new gu-en corpus

English→Kazakh

- Exploit en-kk and kk-ru corpora
- Pivoted back-translation
- Continued training
- Hybridisation with RBMT

All systems scored highly in the official (human) evaluation

Putting it all Together



GoURMET Translate

Input

ગલ્ફ તણાવ : ઈરાનને નહી રોકવામાં આવે તો દુનિયામાં તેલનો ભાવ વધશે - સાઉદી પ્રિન્સ સલમાન

Gujarati

Output

Gulf tensions: If Iran is not stopped, the price of oil in the world will rise - Saudi Prince Salman

English

- Models dockerised with secure API – enables integration of user tools
- First delivery (June 2019):
  - English ↔ Bulgarian, Gujarati, Swahili and Turkish
- Next delivery (March 2020)
  - English ↔ Amharic, Kyrgyz, Serbian and Tamil