gourmet

Global Under-Resourced Media Translation

User Day

Alexandra Birch

18 November, 2021

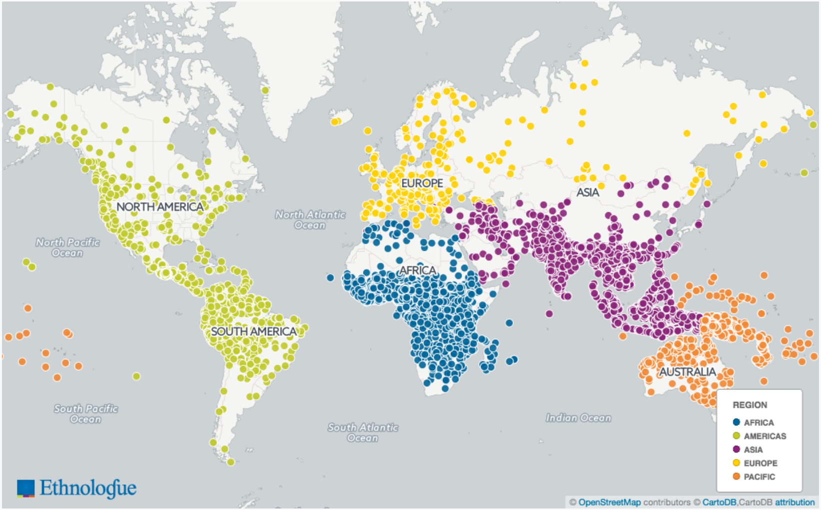
Sometimes You Have Lots of Data ...

Language	**Bitext**	**Monolingual**
Czech	185M	140M
German	571M	237M
Hausa	1.7M	7M
Icelandic	28.2M	101M
Japanese	145.7M	218M
Russian	297M	163M
Chinese	166M	123M
English	—	430M

Facebook 2021 WMT News Task

But what if you don’t have 571M parallel sentences?

Languages in the World



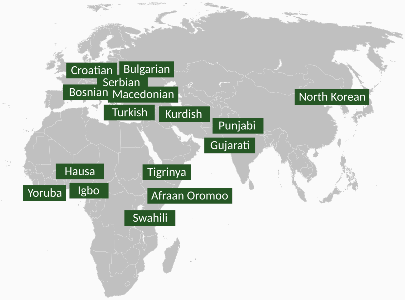
Global Under-Resourced Media Translation

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

THE UNIVERSITY OF EDINBURGH

Universitat d'Alacant
Universidad de Alicante

User Partners

DW

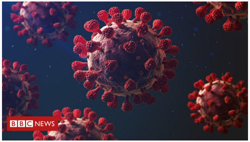
BBC

Use Cases

World Service

Monitoring

Low-resource Domain - Health



Challenges

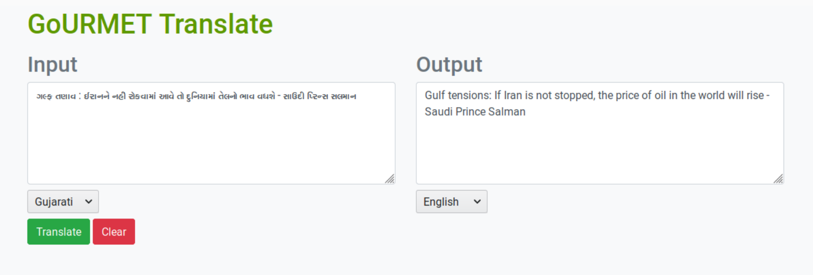
Challenges

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

Approaches

- Data scraping, filtering
- Transfer learning from other languages
- Unsupervised and semi-supervised approaches
- Using linguistic and lexical resources
- Modelling of segmentation and morphology
- Modelling of latent structure

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
- English ↔ Bulgarian, Gujarati, Swahili and Turkish
- English ↔ Amharic, Kyrgyz, Serbian and Tamil
- English ↔ Igbo, Hausa, Macedonian, Tigrinya
- English ↔ Yoruba, Urdu, Turkish, Burmese