Things Needed for Project Proposals

601.764

1/31/23

Cross-Lingual vs. Multilingual

- Cross-Lingual
  - Doing a specific task using methods that span 2 or more languages
- Cross-Language
  - Explicitly crossing a language boundary
  - I.e. Information Retrieval Query in English, but document in Russian
- Multilingual
  - Doing things in multiple languages
  - Not necessarily parallel or in multiple languages at once

All of these are closely related, and the boundaries can be blurry at times

Cross-Lingual Information Extraction

- We want specific information at a sub-sentence level
- Many frameworks and ontologies (likely use multiple)
- Get information from other languages
- I.e. (pun definitely intended), give me Russian cities. More information may be in Russian documents in your corpus

Cross-Language Information Retrieval

- Retrieving documents, not sub-sentence
- Query in language X, document in language Y

Multilingual IR

- IR Systems that can run either in multiple languages independently
- CLIR system that generalizes to more than one language

Multilingual Question Answering

- TyDiQA
- Not cross-lingual?

Multilingual Natural Language Generation

- Generation in multiple languages
- Generation in cross-lingual setting
- Hybrid (aka generation in language X, then project to language Y)

Cross-Lingual NLU

Cross-Lingual Semantics

Cross-Language Summarization

- Similar to generalization, IE, and IR.
- Can you summarize an article in one language in another
- For instance, give me an abstract in Spanish of ACL papers that were written in English

Typology Analysis

- Look at how we can cluster/group/etc. langauges
- Family
- Geography
- Script
- …. Examples here are not very linguistic
- WALS

Phonology Analysis

- How does the audio signal influence methods?

Code-Switching

- Documents or sentences are no longer in only one language
- Data scarcity issues
- Tasks can be a bit simplistic (data issues again)

Multilingual ASR

- Make your ASR system understand more than just English

Multilingual Spoken Language Understanding

- Prime area to be disrupted
- Very little out there
- MASSIVE
- ATIS
- SNIPS

Cross-Lingual Transfer

- Broad, likely many projects will make use of this for something else

Representation Learning

- This is not a Large Language Model course
- But understanding how they work in a multilingual setting is important
- Also has the potential to be a part of multiple projects

Zero-Shot

- Explicitly no labeled data in one language
- How will you evaluate in your project without it?

Few-Shot

- Maybe we have a few examples in some languages?
- How can we leverage this?
- Again, evaluation?
- What IS few-shot?

Low-Resource

- Bibles
- No Language Left Behind (FLORES, FLEURS)
- Low-Resource for a *Specific Task*

Datasets

Popular Tasks

- XNLI
- GLUE
- TyDiQA