Lec 23 Multilingual NLP

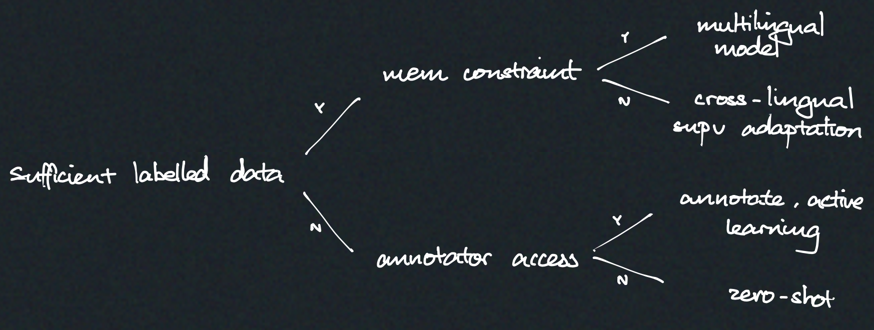
1. Monolingual NLP in multiple langs
2. Cross-lingual NLP

Problems: - paucity of data ... long tail distribution
- not all langs work the same way (linguistically)
  - └ infixes, other scripts, diacritics, ...

Multilingual Learning

Just train a model in multiple langs
+ Easier to deploy
+ Can have transfer learning

Which to pick



Challenges:
- Curse of multilinguality — limited capacity
  - └ per-lang perf ↓ as # lang ↑
- If high resource, sharing param might not be very helpful (when data and compute bottlenecks)
- Tokenisation Disparity
  - └ many byte-level tokens for rare symbols
  - └ doesn't put together semantic chunks
  - └ expensive
  - → Temperature sampling, renormalise vocab frequency by lang resource amount
    - └ can be done when making the vocab or when training
  - → Make vocab size bigger
  - → Heuristic sampling
    - Learn the language sampling weight by evaluating dev set
  - → Train higher resource langs first, then gradually bring in other langs

Machine Translation

Challenges: - Syntax
- Lexical ambiguities

Translation tasks: - WMT shared task
- FLORES — 200 langs translated from wikipedia
- IWSLP (speech)

Models

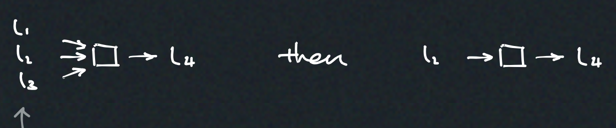
▷ NLLB
- seed data, bitext, monolingual data
- Multilingual embedding model
- Lang ID
- Mixture of experts
- Self supervised (denoising) training
- Back-translation

Multilingual Pre-trained Models

Multilingual Repr Learning

Models
- ▷ mT5
- ▷ mTo
- ▷ byT5
- ▷ Aya

Advanced modelling strats
- ▷ Pre-train then fine-tune



can also be related langs

- ▷ Meta learning
  - └ learn sth good for fine tuning into the target lang
- ▷ Zero-shot transfer
  - ...
- ▷ Annotation Projection
  - └ Induce data with parallel data or bilingual dict
    e.g. project POS tags from Eng in another lang (after alignment)

    Alignment can be done by:
    - Cooccurance states, unsupervised
    - Multilingual BERT then similarity
    - Supervised if have alignment annotations
    - Ask GPT4
- ▷ Picking good lang to transfer from
  ... just by location on globe is good heuristic
- ▷ Normalise different scripts into IPA
- ▷ Sharing params
  - → Just share all
  - → Only share certain part
  - → Generate params per language
  - → Language experts / adaptors
- ▷ Create more data
  - → Active learning : label unlabelled data, filter, and add them to label set

  Want to select representative and uncertain data