11-830 Computational Ethics for NLP

NLP for Good: Lorelei

Carnegie Mellon University
Language Technologies Institute

Government Investment in Languages

- Language Technologies mostly developed for High Resource Languages
  - English, Spanish, German, Arabic, Mandarin
- What about the other 6995 languages?
  - Maybe 30 have good resources (ASR, Treebanks, Parsers)
- What about those around 300-1000?
  - > 1 Millions speakers, Have media (writing systems)
- If no immediate commercial value no support happens

Government Investment in Languages

- Language Technologies mostly developed for High Resource Languages
  - English, Spanish, German, Arabic, Mandarin
- What about the other 6995 languages?
  - Maybe 30 have good resources (ASR, Treebanks, Parsers)
- What about those around 300-1000?
  - > 1 Millions speakers, Have media (writing systems)
- If no immediate commercial value no support happens
- But
  - Wars and Religions!
  - People will spend money to develop non-commercial support if
  - They want to spread the word, (or stop the word)

US Government LT Investment

- DARPA
  - Invested in MT from 1940s
  - Invested in ASR from 1970s
  - Invested in Dialog systems from 1990s
  - Invested in Speech Translation from 1990s
- Case study Lorelei (2015-2020)

The Scenario

- Disaster happens! (e.g. earthquake)
- Area effected doesn't use major language
- Communication is in local language
  - News, TV/Radio, Social Media
- What is going on?
  - Where should you provide support
  - Who is affected
  - How many people need help
  - What is the urgency

Lorelei Incident

- Disaster happens! (e.g. earthquake)
- Communication is in local language
  - News, TV/Radio, Social Media
- Provide
  - Machine Translation
  - NER
  - Situation Frames (11 types) plus location, status, urgency, “gravity”

Lorelei Incident

- Disaster happens! (e.g. earthquake)
- Communication is in local language
  - News, TV/Radio, Social Media
- Provide
  - Machine Translation
  - NER
  - Situation Frames (11 types) plus location, status, urgency, “gravity”
- Do this in
  - 24 hours
  - 7 days
  - 30 days
- You are told the language at hour 0

Lorelei Evaluation Exercises

- May 2016: Dry Run (Mandarin)
- July 2016: Uighur (Turkic Language spoken in Western China)
- July 2017: Tigrinya and Oromo (spoken in Eritrea and Ethiopia)
- July 2018: Kinyarwandan and Sinhala
- Sep 2018: Albanian

Lorelei Performers

- Providing complete systems (with components from elsewhere)
- USC/ISI (with UIUC, Notre Dame)
- CMU (with UW, Melbourne and Leidos)
- BBN (with JHU, UPenn)
- Other components
  - Columbia (urgency, sentiment)
  - UTEP (SF from prosody)

Techniques

- Perform in pronunciation space
  - Not words, morphemes or character space
- Cross Lingual Transfer
  - If w3_l1 co-occurs with w1_l1, w2_l1
  - Maybe w3_l2 means trans(w3_l1) if trans(w1_l1),trans(w2_l2)
  - e.g. China, Japan and Korea vs 中国 , 日本 , 韓国
- Very Low Resources
  - Religious Texts (Bible, Quran and Unix Manuals)
  - Wikipedia
  - Native Informant (“taxi” driver bilingual for limited time)

Techniques

- Global Linguistic Knowledge
  - High morphology language more likely to be free word order
  - Close language borrowing
    - linguistic/geographic/colonial
    - Uighur numbers are Turkish-like
    - Merci is casual Arabic for “thank you”
    - Pashto (Indic) has many Dari/Farsi lexemes
    - “Petrol” might be called “gas”
- Nothing is spelled consistently
  - The dialects aren’t well defined
  - The registers aren’t well defined
  - People code-mix all the time

Lorelei Advances

- Techniques for low resource languages
  - Translation, interpretation, sentiment
  - Both particular languages, and general techniques
- Machine Learning
  - Better use of limited data
  - Not naive just end-to-end
    - Using large mono-lingual dataset to improve models
    - Using structure to make learning easier
- Helping people get immediate help in earthquakes