11-830 Computational Ethics for NLP

Language Technologies for Endangered Languages

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

Language Death

- David Crystal “Language Death” (1977)
- What is an “Endangered Language”?
- What can we do to help save them?
- Should be save them?

Language Death

- Number of Languages
  - How to count
- Names of Languages
  - Own name, other's names
- Language vs Dialect
  - "mutually intelligible"
  - But exceptions: Swedish, Danish, Norwegian
- Half the current languages will die (?)

How Many Speakers to Survive

- Is 500 enough?
  - Depends on community size
  - Depends on community dispersal
  - Depends on community age distribution
- How many is enough?
  - Top 20 languages spoken by 50% of people
- Dutch could become a language for home use only; not for business, education and science

Does Globalization Help

- Language Communities are Distributed
- Communities more exposed to other languages
- Technology encourages global languages
  - Let's use this app to send messages
  - But it doesn't support our languages
  - Font, language, input method, spelling

Different types of Death

- Absorption
  - Code switching, fixed phrases
  - Lexicon continues in plant and place names

What can be done

- An endangered language will progress if its speakers increase their prestige within the dominant community
- An endangered language will progress if its speakers increase their wealth relative to the dominant community
- An endangered language will progress if its speakers increase their legitimate power in the eyes of the dominant community

What can be done

- An endangered language will progress it its speakers have a strong presence in the educational system
- An endangered language will progress if its speakers can write their language down
- An endangered language will progress if its speakers can make use of electronic technology

NLP for Endangered Languages

- Have an on-line representation
  - Unicode method for display
  - Input method (tends towards romanization input)
- Have to accept a standardization
  - English had that too
  - Eth and Thorn ð Þ became th and th
  - Yogh ȝ → z thus Menzies, Dalziel, Calzean

Spelling

- Low resource languages don’t have standard spelling
- Old English texts aren’t standardized
- May take inappropriate writing system
  - e.g. Latin for a Germanic Language
  - Hanzi for Japanese
  - Arabic for Indic language
- Have to merge dialects (or select dialects)
  - English had post-vocalic Rs when it was first written
  - Japanese borrowed English words delete post-vocalic Rs (voice length)
  - “wh” in English became “w” in pronunciation
  - “gh” became something random from X

Spelling Correction

- How many example words do you need to recommend correction?

Spelling Correction

- How many example words do you need to recommend correction?
- Perhaps a few hundred to have > 50% chance of noticing errors
- Take top 500 words
- Build Letter Language model for language
- Given new word:
  - If in 500 its ok
  - If LLM score > threshold accept into list
  - If LLM score < threshold ask if correct
  - Rebuild LLM
  - Have “Teacher” check new words periodically
- A spelling checker for any new language
  - (But codemixing)

Input Method

- Characters often develop for medium
  - Brush strokes for brushed characters (Hanzi)
  - Straight incisions for stone carving (Latin)
  - Triangles for clay tablets (Cuneiform)
- Input method for computers
  - A big keyboard (early Chinese typewriters)
  - A new keyboard (Korean, Japanese)
  - Or just use Romanized input method
  - Or try to teach people a new input method
  - (and they'll use romanized input method)
- Or their writing system will disappear and we'll just use Latin characters

We don’t need no writing system

- Language Technologies for Unwritten Languages
  - Most Languages are not standardly written
  - People may be literate in some other language
- Orality is an interesting thing (Walter Ong)
  - Oral cultures don’t have written memory
  - Speech is the only memory
  - Thus memorable techniques in long stories
  - Rhyming, repetition, alliteration, redundancy and repetition
  - Memory is held in sagas that never change
  - (except they do change)

Nursery Rhymes

- Still part of our oral culture
  - Long term spoken verse
  - Passed down through the ages
  - Rhymes, consistent
  - Though sometimes archaic

Nursery Rhymes

Ring-a-ring o' roses,

A pocket full of posies,

A-tishoo! A-tishoo!

We all fall down.

Nursery Rhymes

Ring-a-round the rosie,

A pocket full of posies,

Ashes! Ashes!

We all fall down

Nursery Rhymes

Sing a song of sixpence,

A pocket full of rye.

Four and twenty blackbirds,

Baked in a pie.

Nursery Rhymes

Half a pound of tuppenny rice,
Half a pound of treacle,
[Mix it up and make it nice, |
  That's the way the money goes]
Pop! goes the weasel.

Nursery Rhymes

- Archaic fixed forms
  - “four and twenty”
  - “posies”
  - “treacle”
  - “daily bread” (Lord's Prayer)
- Archaic Grammar
- Meaning can be obscure

Unwritten Language based Technologies

- Speech based keyword search in arbitrary languages
- Given youtube videos
  - Transcribe them in some generic phonetic form
  - Take keywords from speakers and transcribe them in generic phonetic form
  - Do a match
- Microsoft Research India did this method for low-literate rural farmers
- CMU developed Polly (Rosenfeld et al.)
  - Voice-based job postings with access by keywords

Speech Translation

- Do it from speech not text
- Discover phone-like objects in acoustics
- Find longer segments like “words”
- Have text or speech translation in high-resource language
- Learn standard translation mapping between them
- **Speech** Translation
  - Sitaram et al (CMU, now MSRI) synthesis of unwritten languages
  - Wilkinson et al (CMU, now Amazon) translation of unwritten languages
  - JSALT 2017 Speech/Picture translation for unwritten languages

Endangered Languages

- Language Technologies can help
  - They are only part of the solution
  - More interested in constructing languages than endangered languages
- Should we help?
  - Supporting dying languages will deflect children's competence in major languages (?)
  - Less languages will enable better communication between people (?)
  - What do these languages offer to the world (?)
- Language is culture, identity
  - Denying it is wrong, but what about ignoring it …
- Language diversity is worthy
  - Plant names, disaster warnings