11-737 Multilingual NLP

Code Switching, Pidgins and Creoles

Overview
The multilingual world

• Most people are multilingual
  • These choose the appropriate language for each conversation
• Sometimes they choose multiple languages in the same conversation
  • Generated speech/text contains both languages
  • Not just noun/verb mixing shared grammar/morphology too
• We call this code-switching (or code mixing)
• Monolinguals do this too
  • Register shifting between dialects
  • Colloquial and formal in same utterances
• Sometimes this mix becomes standard
  • Pidgins: non-natives learn a “simplified” mix of multiple languages
  • Creoles: when such a mix/non-native dialect becomes a native languages
**Pidgins and Creoles**

- Creoles have native speakers, Pidgins do not (yet-ish).
- Might be classified as just dialects, possibly low-status
- Jamaican Patois (Creole)
  - An fram Dievid taim op tu wen dem did tek we di Izrel piipl dem an fuos dem fi go wok a Babilan a fuotiin jinarieshan Jiizas did av de-so tu, an fram da taim de tu wen Krais Jiizas baan, a fuotiin jinarieshan dat tu.
- Haitian Creole
  - Zebadya piti Izmayèl la chèf branch fanmi jida a va pi gwo jij pou tou sa ki an rapò ak laalwa peyi a.
- English
  - Is a creole or code-switched Saxon and Norman French (arguably)
Pidgins and Creoles

- These are linguistic terms
- They also get used as political terms too
  - And not always favorably
  - “not proper English”, “dialect”, “uneducated speech”
- Often used for **speech** communication not in writing
  - Cf Latin verses Vulgar Latin
- Thus often hard to find examples, written forms a formal
- But what about code-switching is it a “pre- pidgin”?
Code-switching

- Usually between speakers of equal fluency in two languages
  - Casual speech/text
  - Face-to-face or in social media
- Often defined in terms of a matrix language (based language word order)
  - (major) Word order in matrix languages, as are particles (morphology)
  - Lexemes and short phrases in other language
- We can measure the amount of code switching
  - As percentage (but doesn’t distinguish number of switchers)
  - Other measure try to capture that (but its hard)
Major Code-switching Dialects

● Sometimes between local and global languages
  • But most common examples are between major languages
● Hinglish
  • Hindi and English: very common with educated young Hindi speakers
● Chinglish
  • Chinese and English: common amongst Chinese speakers in Singapore
● Spanglish
  • Spanish and English: common in Spanish speaking areas of US (e.g. South West, New York, Puerto Rico)
● African American English and Standard American English
  • Very common in US in Black communities
● In these cases people are usually very fluent in both languages
When do they code-switch

- Lots of reasons
- Vocabulary coverage:
  - Talk about machine learning in English and food in Hindi
  - Relationships in Spanish, studying in English
- "Showing off" Trends
  - "Fashionable English Works" vs "Ethnic other words"
  - Displaying affiliation (show education and/or show roots)
- Maybe more sentiment in native languages (Rudra et al 2016)
  - Looking at language choice in tweets
- Entrainment (copy others in conversation)
- To get distinctions (simple semantics or opinion)
Why do we care?

- Newspapers and Wikipedia will not be code-switched
  - Why care about NLP of non-standard language?
- Code-switching is how people actually communicate
- Code-switching
  - People type in questions to Google/Bing
  - They talk to call centers
  - They write their opinions
  - Use of code-switching can define group membership
  - People trust code-switched communication better
    - (Not that it is fake, but its their language and someone developed communication in their language)
- Facebook, Amazon, Microsoft, Apple all want to understand Code-switch more
Why is it Hard

- There is very little data available
  - Often code-switched data is removed from datasets
- It is very noisy
  - Spelling is very non-standard so it's hard to know the vocabulary
  - It often romanizes native script (inconsistently)
- Our favorite contextualized word embeddings are confused
  - We have “random” mixed language juxtaposed
  - BERT was never trained for that
  - mBERT was never trained for that
- It's casual speech/text
  - Monolingual casual speech is hard, we now have two languages
Code Switching Data

- Often used by hard to find
  - Harder to verify: is it bilingual or code switched
- Twitter/youtube/reddit
  - Social media is good, but it's not labeled (and very noisy)
- Collecting is hard too
  - Need right environment to have users code-switch
  - Usually based on their peer relationship.
  - There isn’t just one type of code-switching
Code Switching Data

• Annual speech/text workshops on Code-Switching
  • At ACL/Interspeech
• Sitaram et al. “A survey of code-switched speech and language processing”
  • https://arxiv.org/abs/1904.00784
• Thamar Solorio (U Huston)
  • https://ritual.uh.edu/code-switching/
• Growing but still limited
  • Only a few language pairs are studied
Code Switching: LT Tasks

- Language ID (Speech ID)
  - Labeling the words
  - In Speech the pronunciation may vary from monolingual cases
- Speech Recognition/Synthesis
  - Google’s Indian English ASR is actually Hinglish ASR
  - Code-switch synthesis, but ...
- Spelling Normalization
  - As spell checkers don’t work, spelling is particularly inconsistent
  - May require roman → native script conversion too
- POS Tagging
  - Some datasets available (but often noisy)
  - Detecting matrix language can be important
Code Switching: LT Tasks

- Named Entity Recognition
  - (and cross-lingual entity linking)
  - Various challenges have addressed this
- Sentiment analysis
  - Understanding cross-lingual references may be important
- Question Answering
- NLI
- Dialog processing
  - Common Amigos (Ahn et al 2020, Parekh et al 2020)
- When/home to use code-switching
  - Generation as a style
Code Switching: Techniques

- Very similar to general low-resource language issues
  - Find appropriate data
  - Bootstrap labeling
  - Data argumentation/generation techniques
  - Find new (reliable) evaluation techniques
Finding Data

- Social media sites
  - Code Switching (usually) implies casual speech
- Youtube for speech
  - Hinglish: lots of examples, but some human has to find each video
    - Google Indian English ASR can give a reasonable transcription
    - Broadcast news, Bibles/Koran won't have code-switching
- Reddit (or local equivalent)
  - Mining the data is hard
  - You want conversations, not just utterances
- Twitter/Weibo
  - Rarely conversational
  - Sometimes bilingual (translations) not code switching
Bootstrapping Labeling

- Label small amount of data
  - Build classifier for the data
  - Use the classifier to label lots of other data
  - Select “high confidence” samples and add to training data
  - Rebuild classifier
  - Repeat until (something)
- Care has to made to ensure you don’t miss out on important examples
- Care has to made to ensure you don’t just add garbage examples
- Care has to made to ensure you don’t just add trivial addition examples
- Downstream task evaluation would help
  - But you probably don’t have that yet.
Data augmentation/generation

- Build generator from limited data to get more
  - Build classifiers that distinguish real from generated data
  - Choose false positive data to boost base data
  - Been shown to help in lots of cases
  - (could this help building better word embedding models)
- Paraphrase existing data to get more
  - Replace NE, modify some word by translation
Evaluation Techniques

• Task evaluation of held out data
  • Standard techniques, but does that help the overall task
• We don’t yet have lots of good high level tasks to test
  • Dialog understanding
  • Summarization
  • Question/Answering
• MSR India (Khanuja et al ACL 2020)
  • GLUECoS: Set of standard tasks for testing code-switching models
Code-switching, Pidgins and Creoles

• Multilingual is much more varied than one single language
• Code-switching
  • Mixed within an utterance
• Pidgin
  • Non-native mixed lingua franca (often for trade)
• Creole
  • When Pidgin becomes native and its own language
• Issues are as hard as with monolingual casual speech
  • But now we have multiple languages to confuse things
Discussion Points: Code switching

- Select two languages that have some level of code-switching
- We’d like to understand the different **types** of code-switching that happen in that pair or language
- Types might be
  - Matrix language distinctions
  - Cross lingual morphological issue
  - Cross lingual style issues (e.g. politeness, register etc)
  - Low level semantic distinctions
  - Stylistic semantics distinctions (e.g. city vs rural interpretations)
  - Others?