New computational resources for indigenous and minority languages

Kevin Scannell
Saint Louis University
May 13, 2011

Global reach

- Two projects aimed at language revitalization
- Basic goal: allow speakers of indigenous and minority languages to communicate online
- First: Simplified keyboard input
- Second: Community building via Twitter
- Reaching out to under-served languages around the world; the projects support 115 and 90 languages respectively

Extended Latin alphabets

- ~7000 languages in the world; probably less than 1000 have a writing system used with any regularity by the speaker community
- The great majority of languages that do have a writing system employ the Latin alphabet
- Usually augmented by diacritics (å, à, á) or “extended” characters (ŋ, ƙ, ɛ) to better represent the sounds of the language

The Problem

- Extra characters make keyboard input harder
- Rarely physical keyboards, and virtual keyboards not widely installed
- Extra keystrokes required, slower input
- For some languages with emerging written traditions, not all speakers know how to use the extended characters correctly
- Upshot: no one uses them (e.g. 95+% of web texts in Lingala, Yoruba, … omit diacritics)

A Solution: Unicodification

- Unicodification is the process of taking an input text without diacritics or special characters and restoring them automatically

Oll skulu vera fraels at hava sinar askodanir og bera taer fram uttan fordan →

Øll skulu vera fræls at hava sínar áskoðanir og bera tær fram uttan forðan

Ua noa i na kanaka apau ke kuokoa o ka manao a me ka hoike ana i ka manao →

Ua noa i nā kānaka apau ke kūʻokoʻa o ka manaʻo a me ka hōʻike ʻana i ka manaʻo

Moi nguoi deu co quyen tu do ngon luan va bay to quan diem →

Mọi người đều có quyền tự do ngôn luận và bầy tỏ quan điểm

Eni kookan lo ni eto si omi nira lati ni imoran ti o wu u, ki o si so iru imoran bee jade→

Ẹnì kọ̀ọ̀kan ló ní ètọ́ sí òmì nira láti ní ìmọ̀ràn tí ó wù ú, kí ó sì sọ irú ìmọ̀ràn bẹ́ẹ̀ jáde

Machine Learning

- This can be done using statistics, by “training” the computer to recognize where diacritics and special characters belong
- We “study” millions of words of text with the correct special characters for 115 languages
- Training texts are gathered from the web using a web crawler (600+ languages)

Two statistical models

- "Word models" remember words from training and sequences of words to handle ambiguities
- *nios mo na → níos mó ná*
- "Character models" remember sequences of characters: good for morphologically complex languages and those with limited training text
- *riomhchleas fiorchliste → ríomhchleas fíorchliste*

Accentuate.us

- Accentuate.us is a web service created together with my student Michael Schade
- Client programs make requests to our servers, which return unicodified text
- Existing clients: add-on for Mozilla Firefox, the vim text editor, and command-line versions in Perl, Python, Haskell
- All free, open source software
- Demo video...

Using Twitter for language revitalization

- Twitter is a “microblogging” site; users post messages (“tweets”) of 140 characters or less to their “followers”
- I tweet mostly in Irish (@kscanne) and have a couple hundred followers
- Justin Bieber tweets in English (@justinbieber) and has 9.5 million followers
- Free to join; good support for Unicode; users can post from their computer or cell phone; popular among young people

Indigenous Tweets

- Web site that tracks everyone using Twitter in an indigenous or minority language
- Launched St. Patrick's Day 2011 with 35 languages; now supports 90 languages
- Demo: “menu” of people to follow, trending topics per language, sort by any column to help decide who to follow
- Companion blog with interviews of top tweeters every few weeks

Behind the scenes

- Twitter API allows programmers (limited) access to the database of tweets
- Random searches for “statistically unique” words in each language to find candidates
- Statistical language recognition on candidate users' tweets (the hard part)
- If more than a certain threshold are in the language, that user's followers are added to a queue to be checked

Benefits

- Speakers of some endangered languages have started using their language online
- e.g. Gamilaraay (Australia, ~3 speakers), Nawat (El Salvador, ~20 speakers), Delaware (USA, ~78 speakers)
- Engages young people; 6 of the top 10 Irish tweeters are under 30, 3 of 10 under 20
- Competitiveness: top language, top tweeter
- Interesting up-to-the-second corpus data