Saving languages with statistics and web crawlers

BarCamp St. Louis
November 8-9, 2009
Kevin Scannell
Saint Louis University

Language Death

- About 7000 languages spoken in the world
- Estimated that 90% will be extinct by 2100
- Every language is a repository of the culture, traditions, and world view of its community
- The death of a language is an irrevocable loss, comparable to the extinction of a plant or animal species

Saving languages with technology

- Language death is complicated, but a common theme, especially among younger speakers, is that endangered languages aren't suitable for technology, computing, commerce
- Attitudes are reinforced by the reality on the ground: only about 50 languages (0.71%) have fully-localized desktop computer systems
- Firefox 3.5 available in 68 languages (0.97%)
- Spellcheckers for 117 languages (1.67%)

One small piece: keyboard input

- The scripts used to write *most* languages are in Unicode (but not Maldivian, Khamti, ...)
- Yet languages often lack appropriate input methods, or free fonts
- When electronic texts do exist, they are often entered as plain ASCII, either by transliteration (Amharic, ወጣት → weTat), omitting diacritics (Lingala, likɔngá → likonga), or ad hoc approaches (Irish, béal → be/al)

This is terrible

- Diacritics are used to make important distinctions between words (Irish: leite vs.léite, TB 2.0 “Marcáil mar leite”=”Mark as porridge”)
- Texts are invisible to search engines, useless for language modeling
- Literacy issues; speakers must learn two scripts. Or the proper script is forgotten, or never properly learned (tone marks in Kinyarwanda, Lingala)

Web crawling to the rescue

- I have a web crawler at SLU targetting texts in endangered languages; 431 at present
- See *http://borel.slu.edu/crubadan/*
- Volunteers from around the world are editing frequency lists from these text collections to generate new open source spell checkers (more than 20 new ones to date)

Diacritic restoration script

“charlifter”: takes as input plain ASCII text in some language, and outputs the same text in its proper Unicode rendering:

```
$ echo "an chead teanga oifigiuil" | sf.pl -r ga

 an chéad teanga oifigiúil

$ echo "Ngolo, nina, zambi ikamwisi bango." | sf.pl
-r ln

Ngɔlɔ, niná, zambí ikamwísí bangó.

$ echo "Olelo aku 'o Papa" | sf.pl -r haw

ʻŌlelo aku ʻo Pāpā

$ echo "My tu bo ke hoach la chan ten lua" | sf.pl
-r vi

Mỹ từ bỏ kế hoạch lá chắn tên lửa
```

Language-independent, trainable

- No knowledge of the language is needed to train an new language model
- Just feed it enough properly-encoded text and the engine “learns” how to restore texts
- Existing models trained with web-crawled texts
- Bayesian learner, with two kinds of features: word-internal features (three character sequences), and word sequences when there are real ambiguities (leite/léite).

Applications

- End-user applications: whenever text input is needed: word processing, texting on phones
- Statistical modeling: 98% of web text in Lingala is ASCII, but trained charlifter on the 2% and restored the 98%. Provided large enough frequency lists to generate a good spell checker in proper encoding.
- Help wanted: Firefox addons, OpenOffice extensions, web app, phones?