Accentuate Us!

Kevin Scannell and Michael Schade

November 10, 2010

Language Death

- About 7000 languages spoken in the world
- More than 90% are expected to disappear before 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
- Many endangered language communities are looking to the internet and technology to help revitalize their language

Endangered Language Technology

- 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%)
- For this talk, we're looking at something even more basic: keyboard input

Keyboard Input

- The majority of languages are oral - no written tradition
- Good news: among those that have writing systems, almost all scripts are available in Unicode (but not Maldivian, Khamti, ...)
- Yet even Unicode-encoded languages often lack appropriate input methods, or free fonts
- When electronic texts do exist, they are often entered as plain ASCII, either by transliteration (Cherokee, □□□□□ → galvquodiyu), omitting diacritics (Lingala, likɔngá → likonga), or ad hoc approaches (Irish, béal → be/al)
- Omitted diacritics matter!   Leads to ambiguities, misunderstandings (leite vs. léite).

Diacritic Restoration

- Software that takes plain ASCII text in some language as input, and outputs the text with all diacritics or extended characters in place
- Examples
- 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
- Uwe setin suen gha gu emwa ni sike rue ghae emwi esi ne uwe rue →
  Uwẹ sẹtin suẹn gha gu emwa ni sikẹ ruẹ ghae emwi esi ne uwẹ ruẹ
- 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
- Tout moun gen dwa a libete lide yo ak lapawol yo →
  Tout moun gen dwa a libète lide yo ak lapawòl yo
- 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 kòọ̀ḳan ló ní èṭó ̣sí òmì nira láti ní ìmòṛàn tí ó wù ú, kí ó sì sọ irú ìmòṛàn béẹ̀ jáde

Statistical Machine Learning

- Given an ASCII input, every character that allows a diacritic or an extended form represents a "classification problem"
- We use a machine learning approach; the program learns where the diacritics belong by gathering statistics from a "training corpus" of texts with the diacritics in place
- Remembers words seen in training data; statistics on co-occurring words to deal with ambiguous cases (Irish "leite" vs. "léite" or even English "resume" vs. "résumé")
- For never-before seen words, uses statistics of 3-character sequences in a neighborhood of the character in question (French initial "cera" vs. "cerc", "cabl" vs. "cabo"). This is the generic case for under-resourced languages
- Training texts crawled from the web; 114 in all!

API

- Protocol: JSON
- Calls
  - langs
  - lift
  - feedback
- Sample Call

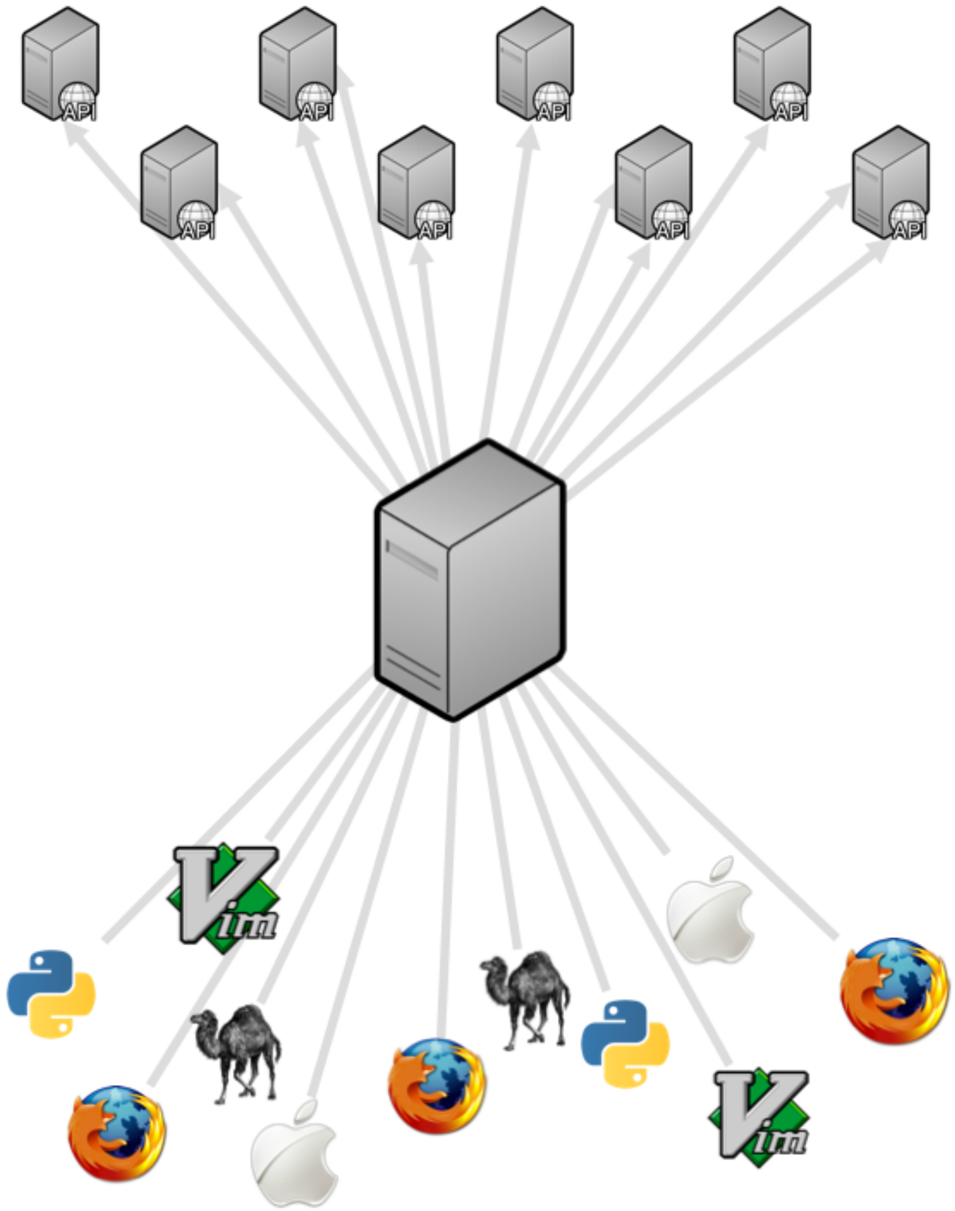
  -
    ```
    {   "call": "charlifter.lift"
      , "lang": "ht"
      , "text": "Bon, la fe sa apre demen pito, le la we mwen andey."
      , "locale": "ht"
    }
    ```
- Full documentation at http://accentuate.us/api

Service Architecture

Geographically Distributed
API Servers

Load-Balancing Proxy

Clients



HTTP Communication (Proxy)

```
Cache-Control: no-cache
Connection: keep-alive
Pragma: no-cache
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7
Accept-Encoding: gzip,deflate
Accept-Language: en-us,en;q=0.5
Host: ht.api.accentuate.us:8080
User-Agent: Accentuate.us/0.9b3 Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US; rv:1.9.2.10) Gecko/20100914 Firefox/3.6.1
Content-Length: 113
Content-Type: application/json; charset=utf-8
Keep-Alive: 115

{"call":"charlifter.lift","lang":"ht","text":"Bon, la fe sa apre demen pito, le la we mwen andey.","locale":"ht"}
```

HTTP Communication (API)

```
Cache-Control: no-cache
Connection: close
Pragma: no-cache
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7
Accept-Encoding: gzip,deflate
Accept-Language: en-us,en;q=0.5
Host: ht
User-Agent: Accentuate.us/distribution
Content-Length: 113
Content-Type: application/json; charset=utf-8

{"call":"charlifter.lift","lang":"ht","text":"Bon, la fe sa apre demen pito, le la
we mwen andey.","locale":"ht"}
```

Clients

- Perl

  ```
  $ ./sf-client.pl -r -l ga -i "Is i an Ghaeilge an chead teanga oifigiuil."
  Is í an Ghaeilge an chéad teanga oifigiúil.
  ```

- Python
- Vim
- OS X Service
- OpenOffice.org

Mozilla Firefox

- UI and Localization Decisions
- Implement API calls
    - Langs
        - Silent
    - Feedback
        - Opt-in
        - Improve language models
    - Lift
        - Complicated!
- Looking ahead

Demos

accentuate.us

Thank You!