facebook

facebook

Social (distributed) language modeling, clustering and dialectometry

David Ellis
August 7, 2009
TextGraphs (@ACL)

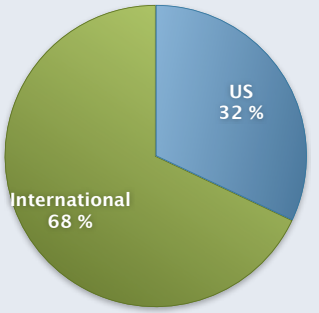
Why?

- Users
  - connect w/ wider audience
  - access to more info (search)
- Facebook
  - growth, open communication
- Personal
  - Google (mach.) => FB (crowd)
- Research
  - corpora (public, anonymized)
  - tools (API, open source)

Sharing with the world

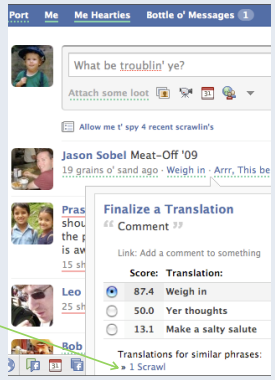
- Facebook is live in 64 languages
- 55 more are in translation (مال شما؟)
- Over 2/3 of users are outside the US
- >50% of Internet users in Chile are active Facebook members
- Help 350,000 apps go global

- Share docs (social productivity)
- Local insight => travel info
- Explore common interests
- Play games (MasMultilingualOn)
- Taste virtual cuisine (and learn to cook the real kind)



Translations on Facebook

- Community-driven, open
  - glossary => inline (voting) => local UI
- French <24h
- English (Pirate)
  - ~30,000 users
  - >50,000 (phrase) translations submitted
- Similar suggestions
  - fuzzy clustering, regional sharing, MT*
- आप दुनिया चैट मदद कर सकता है!



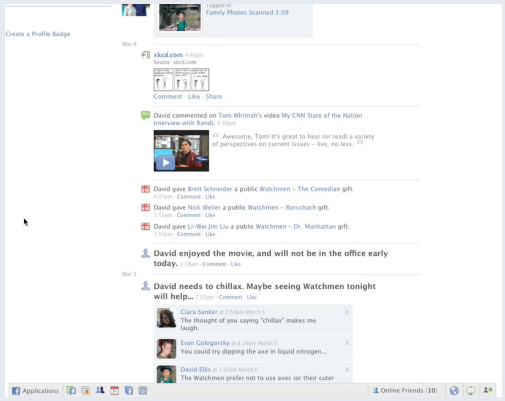
*experiments with open source toolkits in progress

Dynamic text as {token}s

- {name} depends on
  - person
  - context
- food pyramid app
  - "Eat a(n) {fruit}!"
  - kiwi? apple?
- translations
  - "Syö(kää) {fruit}!"
  - sing/plural?
    - addressee (of imperative)

Linguistic tools

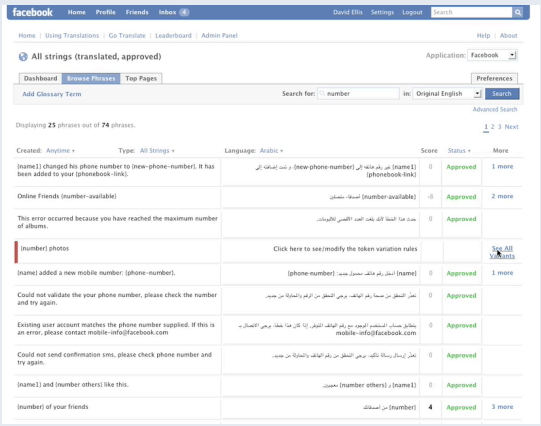
- Rules
  - orthographic
  - phonological
- Features
  - gender
    - Català
  - number (si, du, pl)
    - Russian, Arabic
- Models
  - language
  - pronunciation



Linguistic tools

- Variation
  - native explosion
    - 1-to-many
  - cluster folding
    - many-to-1
  - detection
    - spam (text)
    - botnet (usage)
    - authentic?

- "Güd [pet]tRn'd!" -EG



Deep localization

- Transform layout (not just text)

أرحب في أعماق روحك...
قبل أن يغادر ، انقر هنا
لالأقران في منجم
هل لي أن تنضم في
المستقبل (في سنغافورة ،
كاليفورنيا وراء ذلك الوقت)؟

Fu(ture|n)

- Open innovation
  - experimental API
    - access [via Hive] to publicly shared, anonymized data
      - social graph
      - language model
  - parallel (dynamic) corpus
    - named entity annotated
    - growing, with snapshots
- Instant, free multimodal interpretation
  - http://www.colips.org/blog/acl-ijcnlp-2009/index.php/lieutenant-commander-data-star-trek/comment-page-1/#comment-292

Loc
||
˅
Fuzz
||
˅
SMT
||
˅
?

Questions?
Kiitti kaikille!

facebook