Textual Corpora, Treebanks, and the Human Language Project

Steven Abney

Department of Linguistics
Computer Science and Engineering
School of Information

2015 Mar 30
What kinds of (annotated) text corpora do computational linguists use?

A research example: using existing treebanks to bootstrap treebanks in other languages

Extrapolating into the future: building a comprehensive multilingual dataset – a Human Genome Project for language

What a library might provide
Textual Data

Plain text

- Language samples
  - Plain text
  - Preferably large amounts (1M–1B words)
  - In as many languages as possible

- Some LDC titles:
  - UN Parallel Text (Complete) [1994]
  - European Language Newspaper Text [1995]
  - Spanish News Text [1995]
  - Mandarin Chinese News Text [1995]
  - CALLHOME Egyptian Arabic Transcripts [1997]
  - Portuguese Newswire Text [1999]
  - Korean Newswire [2000]
  - HUB5 German Transcripts [2003]
  - Chinese Gigaword [2003]
  - Arabic Gigaword [2003]
  - Czech Broadcast News Transcripts [2004]
  - Web 1T 5-gram Version 1 [2006]
  - Hungarian-English Parallel Text, Version 1.0 [2008]
  - Web 1T 5-gram, 10 European Languages Version 1 [2009]

- Aligned texts, multiple translations
Textual data
Lexicons

- Wordnet: lexical database
  - hypernymy, synonymy, meronymy
  - “Synset” = set of words that share a meaning
- Multi-lingual wordnets – synsets that cross languages (translation equivalents)
- Babelnet – combines Wordnets with Wikipedia and Wiktionary
- Panlex – database constructed from thousands of bilingual print dictionaries
Textual data
Parts of speech

- the Brown corpus:
- Why?
  - Input for machine learning
  - Automatically train a system to label new text
  - First step in language-interpretation pipeline

The Fulton County Grand Jury said Friday an investigation of Atlanta’s recent primary election produced “no evidence” that any.

- at
- np-tl
- nn-tl
- jj-tl
- nn-tl
- vbd
- nr
- at
- nn
- in
- np$
- jj
- nn
- nn
- vbd
- “
- “
- at
- nn
- ”
- cs
- dti
El portavoz del Consejo Político Municipal de IU en Santander, Ernesto Gómez de la Hera, explicó hoy en conferencia de prensa que este ...
Textual data
Constituent-structure treebanks

- Pierre Vinken, 61 years old, will join the board as a nonexecutive director in Nov. 29.

Penn Treebank
Textual data
Dependency treebanks

- Much more compact than constituent trees, equivalent for practical purposes

- Purpose: training a parser (interpretation, translation)
Research Example

How can we learn a parser without a treebank?

Motivations

- Linguistics: ultimate subject matter is human language capacity = ability to learn language.
- Google: access to data in all languages
- DARPA: decision support in crisis management; information extraction from news media and social media

Treebanks

- I know of treebanks for 43 languages:

  Arabic, Armenian, Ancient Basque Bulgarian Catalan Chinese Czech Danish Dutch English English, Early Modern English, Middle English, Old Estonian Finnish French German Gothic Greek Greek, Ancient Hebrew Hindi-Urdu Hungarian Icelandic Indonesian Italian Japanese Karuk Korean Latvian Lithuanian Malagasy Maltese Manx Maori Mandarin Manx Marathi Macedonian Malay Malayalam Manx Mapudungun Maori Marathi Macedonian Malay Malayalam Manx

But there are 6800 languages (Ethnologue)
Research Example
Learning a dependency parser for a new language

- Accuracy measure: percentage of governors correctly identified

  - Monolingual grammatical inference: 47%
  - Delexicalized transfer: 52%
  - Multi-source delexicalized transfer: 55%
  - Adaptation using bitexts: 59%
  - Adaptation using bitexts + language relationships: 62%
  - Supervised training: 84%

_McDonald et al 2012_

- Existing methods neglect bilingual dictionaries
Research Example
A challenge: low-resource languages

- Making the methods practical for **low-resource languages**
  - Well-resourced languages: $\sim 50 = 0.7\%$
  - E.g., Google translates 57 languages
  - All but 18 are Indo-european, none are endangered.

- Example: machine translation
  - Current methods require 2–10 million words of **bitext** for training
  - Largest source of bitext is the Bible: 0.8 M words, 459 languages (7%).
  - New Testament: 0.1 M words, 1213 languages (18%)
The Human Language Project
Bootstrapping resources for low-resource languages

- Where we would really like to go
  - Comprehensive language resources
  - The Human Genome Project for languages

- Made urgent by language endangerment
  - “Low resource” = digitally endangered
  - 33% endangered, another 10% vulnerable
  - Half of the world’s languages have fewer than 6,000 speakers.
  - 4% have gone extinct since 1950
  - Current rate of extinction: 2 languages/month
  - Projections: 50–90% loss by end of century
The Human Language Project

Resources

What do we mean by resources?
- Target-language plain text
- Monolingual dictionaries with parts of speech
- Bilingual dictionaries
- Morphological paradigms
- Bitext
- Treebank

All expressible in a simple data format
What Libraries Might Provide

- Scanned books

- Traditional language description
  - A grammar and a lexicon, maybe a text collection
  - Printed books, for human consumption
  - All that is available for most languages

---

Dat.  Ako-then, theck, sen, seck, to themselves.
Acc.  Ako, themselves.
Abl.  Ako-khon, khanak, from themselves.
Loc.  Ako-re, talare, in, on themselves.

Lars Skrefsrud, *A grammar of the Santhal language* (1873)

---

What Libraries Might Provide
Scanned books

- **Difficulties**
  - OCR is a huge problem – poor or nonexistent for non-roman scripts, poor for text with diacritics
  - Possible alternative: crowd-sourcing
  - Transcription is one thing, conversion to dataset is another
What Libraries Might Provide

Libraries as digital archives?

- Biggest current provider: Linguistic Data Consortium
  - Heavy emphasis on speech, languages with overwhelming commercial and intelligence value (English, Chinese, Arabic, western Europe)
  - Expensive
- Language documentation archives
  - Archiving of traditional field notes, recordings
  - Small scale, little support for or awareness of computational methods
  - Access is often highly restricted
What Libraries Might Provide
Libraries as digital archives?

- What is lacking: archives that provide
  - Free public access
  - Comprehensiveness
  - Machine consumption

- Broader issues: incentives/hindrances to producing resources
  - Recognition of dataset production as publication
  - Production of machine-oriented datasets from copyrighted print works
  - Ability to publish annotation of others’ datasets