Automatic Extraction of Linguistic Data from Digitized Documents

Terrence Szymanski
tdszyman@umich.edu

Berkeley Linguistics Society 39

February 17, 2013
1 Motivation, Goals

2 Language ID

3 Translation ID

4 Performance

5 Future Directions
John Goldsmith (2007) *A New Empiricism*

“[T]he goal of the linguist is to provide the most compact overall description of *all of the linguistic data that exists at present*”

– John Goldsmith

Steven Abney (2011) *Data-Intensive Experimental Linguistics*

“[A]ny experimental foray into universal linguistics will be a data-intensive undertaking. It will require substantial samples of many languages—ultimately all human languages—in a consistent form that supports automated processing across languages.”

– Steven Abney
Motivation

The long view

1. The goal of universal linguistics is to explain structures of all human languages.
2. Rigorous, large-scale analysis is best done with help of a computer.
3. Therefore, we need computer-readable data from all languages.

The short view

1. Let’s start with the data that’s available.
2. Many digital resources aren’t machine readable.
Sources of Machine-Readable Linguistic Data

Currently available
- NLP corpora
- PDFs of linguistics papers, via ODIN (Lewis & Xia, 2010)
  - [odin.linguistlist.org](http://odin.linguistlist.org)

Currently unavailable
- Undocumented languages
- Field notes and unpublished material
- Non-digitized material
- Unstructured digital material
  - e.g. Digitized books in online libraries
Availability of Language Data

Language Coverage of Current Digital Resources

<table>
<thead>
<tr>
<th>Number of Languages Included</th>
<th>Fraction of World Languages Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treebanks</td>
<td>0.5%</td>
</tr>
<tr>
<td>Tatoeba</td>
<td>1.3%</td>
</tr>
<tr>
<td>Google Translate '06</td>
<td>0.9%</td>
</tr>
<tr>
<td>Google Translate '07</td>
<td>4.1%</td>
</tr>
<tr>
<td>Google Translate '08</td>
<td></td>
</tr>
<tr>
<td>Google Translate '09</td>
<td></td>
</tr>
<tr>
<td>Google Translate '10</td>
<td></td>
</tr>
<tr>
<td>Google Translate '11</td>
<td></td>
</tr>
<tr>
<td>Google Translate '12</td>
<td></td>
</tr>
<tr>
<td>Wikipedia '01</td>
<td></td>
</tr>
<tr>
<td>Wikipedia '02</td>
<td></td>
</tr>
<tr>
<td>Wikipedia '03</td>
<td></td>
</tr>
<tr>
<td>Wikipedia '04</td>
<td></td>
</tr>
<tr>
<td>Wikipedia '05</td>
<td></td>
</tr>
<tr>
<td>Wikipedia '06</td>
<td></td>
</tr>
<tr>
<td>Wikipedia '07</td>
<td></td>
</tr>
<tr>
<td>Wikipedia '08</td>
<td></td>
</tr>
<tr>
<td>Wikipedia '09</td>
<td></td>
</tr>
<tr>
<td>Wikipedia '10</td>
<td></td>
</tr>
<tr>
<td>Wikipedia '11</td>
<td></td>
</tr>
<tr>
<td>Wikipedia '12</td>
<td></td>
</tr>
</tbody>
</table>

Automatic Extraction of Digital Linguistic Data

Terrence Szymanski tdszyman@umich.edu
Digitization projects (The Hathi Trust, Google Books, Project Gutenberg, et al.) include millions of books. *Some* of those books contain language data valuable to linguists, e.g:

- Grammars (e.g. *A Grammar of the Santhdal Language*)
- Lexicons (e.g. *Trukese-English Dictionary*)
- Readers and texts (bilingual or monolingual) (e.g. *Kickapoo Tales*)

**Challenges**

- OCR (optical character recognition) is weak.
- Some texts are subject to copyright restrictions.
- Quality of data is uncertain.
Motivation, Goals

Desired Input and Output

Electronic Document

Parallel Corpus (Bitext)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>F-52</td>
<td>holako hechlen, onkodo okaena?</td>
</tr>
<tr>
<td>E-52</td>
<td>they who came yesterday, what has become of them?</td>
</tr>
<tr>
<td>F-53</td>
<td>Hopon em ranade tae, oni joharam lagit’e hechakana</td>
</tr>
<tr>
<td>E-53</td>
<td>whose son you gave medicine to, he has come to thank you</td>
</tr>
<tr>
<td>F-54</td>
<td>Enbetarem ranade, oni do phariaoena,</td>
</tr>
<tr>
<td>E-54</td>
<td>to whom you gave medicine at that time, he has recovered.</td>
</tr>
</tbody>
</table>

Figure: The high-level objective of bitext data collection.
Language ID

**Language ID Task**

**Input:** a multilingual electronic text.

**Output:** language tag for each token in the text.

- Assume one of the languages is well-known (e.g. English), and
- The other language is unknown (i.e. no text available to train a language model).
- Dictionary approach is problematic given OCR text.
- **Unsupervised** approach: ID English and non-English.
- **Semi-Supervised** approach: manually tag a small number of non-English tokens.
The Santal language has no special relative pronouns for animate objects, but supplies their place by the personal (full and abridged), the interrogative, and demonstrative pronouns; as holako hecklen, onkodo okaena? they who came yesterday, what has become of them? what has become of them who came yesterday? Hopen em rānade tae, oni johar-am lagit’e hechkana, whose son you gave medicine to, he has come to thank you. Enbetarem rānade, oni do phariaoena,

from *A grammar of the Santhal language* by L. O. Skrefsrud, 1873.
The Santal language has no special relative pronouns for animate objects, but supplies their place by the personal (full and abridged), the interrogative, and demonstrative pronouns; as "holuko hecklen, onkodo okaena?" they who came yesterday, what has become of them? what has become of them who came yesterday? "Hopon em rānade tae, oni johar- nam lagitē hechakana, whose son you gave medicine to, he has come to thank you. "Enbetarem rānade, oni do phariaoena,

from A grammar of the Santhal language by L. O. Skresrud, 1873.
Word-level Language ID

English vs. Known

- Supervised: requires some amount of labeled data.
- Train a Support Vector Machine using n-gram features.
- Evaluated using 2,600 hand-annotated tokens from the Santhul text: 82% Precision, 66% Recall.

English vs. Unknown

- Train an n-gram model of English.
- Estimate a single decision boundary using known non-English text. This boundary is then used to classify all languages: no language-specific labeled data is needed.
- Evaluated on English vs. Dutch/German. (c. 50k words): 86% accuracy.
English vs. Other Language Modeling

5-gram Character Model of English, Word LogProb

- English
- German
- Dutch
Translation Identification

**Translation ID Task**

**Input:** a multilingual text with spans of foreign text identified.

**Output:** for each foreign text span, a span of English text representing a translation of the foreign text.

- Assume that the English translation immediately precedes or follows the foreign text.
- Assume the length of the translation is roughly the same length (in characters) as the foreign text.
- Use statistical word alignments to choose the better candidate translation.
The Santal language has no special relative pronouns for animate objects, but supplies their place by the personal (full and abridged), the interrogative, and demonstrative pronouns; as holoko hecklen, onkodo okaena? they who came yesterday, what has become of them? what has become of them who came yesterday? Hopon em rānade tae, oni johar-am lagite hechakana, whose son you gave medicine to, he has come to thank you. Enbetarem rānade, oni do phariaoena,
The Santal language has no special relative pronouns for animate objects, but supplies their place by the personal (full and abridged), the interrogative, and demonstrative pronouns; as 'holuko hecklen, onkodo okaena?' they who came yesterday, what has become of them? what has become of them who came yesterday? 'Hopon em rānade tae, oni johar-am lagit'e hechakana, whose son you gave medicine to, he has come to thank you. 'Enbetarem rānade, oni do phuriaoea,'
Translation Selection Experiment

<table>
<thead>
<tr>
<th>Sentence and Candidate Translations</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;He abused our trust.&quot;</td>
<td></td>
</tr>
<tr>
<td>a) Il a abusé de notre confiance.</td>
<td>18.5</td>
</tr>
<tr>
<td>b) Il éclata en larmes.</td>
<td>40.3</td>
</tr>
<tr>
<td>&quot;The floor was covered with blood.&quot;</td>
<td></td>
</tr>
<tr>
<td>a) Le sol était couvert de sang.</td>
<td>15.9</td>
</tr>
<tr>
<td>b) La machine était recouverte de poussière.</td>
<td>46.7</td>
</tr>
</tbody>
</table>

- Each sentence is paired with two candidate translations.
- Translation model (GIZA++) is trained on all pairs (50% noise).
- The model assigns an alignment cost to each sentence pair.
- The lower-cost translation is chosen as correct.
- Accuracy:
  - 500 sentences 73%
  - 5k sentences 88%
  - 50k sentences 94%
Performance and Evaluation

How does this process fare on actual OCR e-books?

A Grammar of the Santhali Language (Skrefsrud, 1873)

- 389 pages (190k word tokens).
- 15 annotated pages (7k word tokens).
- Use annotated pages to train SVM language ID classifier.
- Consider all sequences of 2+ foreign words as potential bitexts.

- Estimating recall is problematic.
- Sample 100 predicted bitexts for evaluation:
  - **99%** correct foreign language ID (precision)
  - Of these 99, 69 have adjacent translations
  - Of these 69, 19 (**28%**) had the translation approximately correctly identified.

- Room for improvement (following slides).
Examples of bitext predictions from the Santhali grammar. (Foreign text in bold; predicted gloss underlined.)

1. had struck him. had struck him. he had struck hitn.
   DUAL. DUAL. DUAL.
   I D-al-a1, kat'-ti; 4-ta- Dal-akat'-li.-tc-th' - Paset'-e-dat-a-cat'-liti..
   lt-1can-a-e, He kan-A-han-e, If tcth-loan, Perhaps
   had struck us he had struck us he had struck us
   strike.
   INCHOATIVE PAST.
   Dal-Jko-dagidoll-kan-tahVkan,
   Tlhey whom they were about
   to strike.
   OPTATIVE.

2. oni hola-m del-led-e, what has become of him whom you
   saw yesterday? This is much more elegant and certainly more
   correct than to say: oni hola-m diel-ed-e-a, oni do okare,
   for the latter means literally: you saw him yesterday, what
   has become of him?

3. 

Automatic Extraction of Digital Linguistic Data
Terrence Szymanski tdszyman@umich.edu
OCR Troubles

Even if the bitexts are extracted perfectly, OCR errors limit their utility for further processing.

- OCR has trouble with diacritic marks.
- Layout and font information is lost.
- Using different OCR software could help.
Is this line of work worth continuing?

1. Is the objective (machine-readable data from all languages) worthwhile?
2. Is this approach to data collection the right one?

- Is OCR text too noisy to be useful?
- Are automated approaches more useful than manual (e.g. crowd-sourcing)?
- Better models for language ID?
- Better models for gloss detection?
Future Directions

Is this line of work worth continuing?

1. Is the objective (machine-readable data from all languages) worthwhile?
2. Is this approach to data collection the right one?

- Is OCR text too noisy to be useful? (Maybe, not necessarily)
- Are automated approaches more useful than manual (e.g. crowd-sourcing)? (Need a mix)
- Better models for language ID? (see next slide)
- Better models for gloss detection? (see next slide)
Future Directions

Better models for language ID?
- Incorporate typographic features (where available)
- Better models of page layout (i.e. tables, lists)
- Sequential models for language ID (e.g. hierarchical HMMs)

Better models for gloss detection?
- Automatically determine translation length
- Incorporate typographic features and page layout
- Look at cue phrases such as “which means” that indicate translations.

Reducing OCR errors
- Commercial OCR software seems to fix many errors.
- There is no good language-agnostic OCR software.
References

S. Abney.
Data-intensive experimental linguistics.

J. Goldsmith.
Towards a new empiricism.

C. Kruengkrai, P. Srichaivattana, V. Sornlertlamvanich, and H. Isahara.
Language identification based on string kernels.
KICSS. 2006.

W. D. Lewis and F. Xia.
Developing ODIN: A multilingual repository of annotated language data for hundreds of the world’s languages.

F. J. Och and H. Ney.
A systematic comparison of various statistical alignment models.

L. O. Skrefsrud.
A grammar of the Santhal language.
1873.
Thank You

Questions?

Terry Szymanski

tdszyman@umich.edu
www-personal.umich.edu/~tdszyman/

Thanks to Steven Abney, Ezra Keshet, and to the Google Digital Humanities Awards Program for partially supporting this work.