Corpus Acquisition from the Interwebs

Christian Buck,
University of Edinburgh
“There is no data like more data”
(Bob Mercer, 1985)
Mining Bilingual Text

"Same text in different languages"

- Usually: one side translation of the other
- Full page or interface/content only
- Potentially translation from same page
  - Twitter, Facebook posts
- Human translation preferred
- Anything that improves BLEU

Structural Translation Recognition, Acquiring Natural Data
Pipeline

1. Candidate Generation
2. Candidate Ranking
3. Filtering
4. Optional: Sentence Alignment
5. Evaluation
STRAND: parent pages

A page that links to different language versions

English  French  Spanish

x.com/en/cat.html
x.com/fr/chat.html

Require that links are close together
Example parent page
STRAND: sibling pages

A page that links to itself in another language
Candidate Generation without links

1. Find and download multilingual sites
2. Find some URL pattern to generate candidate pairs

<table>
<thead>
<tr>
<th>xyz.com/en/</th>
<th>xyz.com/fr/</th>
</tr>
</thead>
<tbody>
<tr>
<td>xyz.com/bla.htm</td>
<td>xyz.com/bla.htm?lang=FR</td>
</tr>
<tr>
<td>xyz.com/the_cat</td>
<td>xyz.fr/le_chat</td>
</tr>
</tbody>
</table>
Generating language independent URLs

http://100.steelcase.com/de/2012/03/
http://100.steelcase.com/fr/2012/03/
http://100.steelcase.com/es/2012/03/

Common key:
http://100.steelcase.com//2012/03/
<table>
<thead>
<tr>
<th>Count</th>
<th>Parameter</th>
<th>Count</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>545875</td>
<td>lang=en</td>
<td>33503</td>
<td>lang=eng</td>
</tr>
<tr>
<td>140420</td>
<td>lng=en</td>
<td>19421</td>
<td>uil=English</td>
</tr>
<tr>
<td>126434</td>
<td>LANG=en</td>
<td>15170</td>
<td>ln=en</td>
</tr>
<tr>
<td>110639</td>
<td>hl=en</td>
<td>14242</td>
<td>Language=EN</td>
</tr>
<tr>
<td>99065</td>
<td>language=en</td>
<td>13948</td>
<td>lang=EN</td>
</tr>
<tr>
<td>81471</td>
<td>tlng=en</td>
<td>12108</td>
<td>language=english</td>
</tr>
<tr>
<td>56968</td>
<td>l=en</td>
<td>11997</td>
<td>lang=engcro</td>
</tr>
<tr>
<td>47504</td>
<td>locale=en</td>
<td>11646</td>
<td>store=en</td>
</tr>
<tr>
<td>33656</td>
<td>langue=en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>Language</td>
<td>Value</td>
<td>Setting</td>
</tr>
<tr>
<td>--------</td>
<td>----------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td>545875</td>
<td>lang=en</td>
<td>130610</td>
<td>setlang=pt</td>
</tr>
<tr>
<td>479973</td>
<td>lang=fr</td>
<td>127735</td>
<td>ang=nl</td>
</tr>
<tr>
<td>326970</td>
<td>lang=de</td>
<td>126434</td>
<td>LANG=en</td>
</tr>
<tr>
<td>290010</td>
<td>lang=es</td>
<td>106438</td>
<td>setlang=de</td>
</tr>
<tr>
<td>171125</td>
<td>setlang=fr</td>
<td>104836</td>
<td>setlang=es</td>
</tr>
<tr>
<td>158602</td>
<td>lang=it</td>
<td>99065</td>
<td>language=en</td>
</tr>
<tr>
<td>152016</td>
<td>langue=fr</td>
<td>95876</td>
<td>language=fr</td>
</tr>
<tr>
<td>151220</td>
<td>setlang=it</td>
<td>83050</td>
<td>language=es</td>
</tr>
<tr>
<td>142964</td>
<td>setlang=sv</td>
<td></td>
<td></td>
</tr>
<tr>
<td>131258</td>
<td>language=de</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Filtering Candidates: Length

Extract texts and compare lengths (Smith 2001)

\[ \text{Length}(E) \approx C \times \text{Length}(F) \]

learned, language-specific parameter

Document- or sentence-level
Filtering Candidates: Structure

<html>
<body>
<h1>Where is the cat?</h1>
The cat sat on the mat.
</body>
</html>
Filtering Candidates: Structure

<html>
<body>
<h1>Where is the cat?</h1>
The cat sat on the mat.
</body>
</html>

<html>
<body>
El gato se sentó en la alfombra.
</body>
</html>
Linearized Structure

[Start:html]
[Start:body]
[Start:h1]
[Chunk:17 Bytes]
[End:h1]
[Chunk:23bytes]
[End:body]
[End:html]
Levenshtein Alignment

[Start:html] Keep
[Start:body] Keep
[Start:h1] Delete
[Chunk:17 Bytes] Delete
[End:h1] Delete
[Chunk:23 Bytes] 23 Bytes -> 32 Bytes
[End:body] Keep
[End:html] Keep
Levenshtein Alignment

[Start:html]  Keep
[Start:body]  Keep
[Start:h1]   Delete
[Chunk:17 Bytes]  Delete
[End:h1]  Delete
[Chunk:23 Bytes]  23 Bytes -> 32 Bytes
[End:body]  Keep
[End:html]  Keep

When does this break?
Variables characterizing alignment quality

<table>
<thead>
<tr>
<th>dp</th>
<th>% inserted/deleted tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td># aligned text chunks of unequal length</td>
</tr>
<tr>
<td>r</td>
<td>(Pearson) correlation of lengths of aligned text chunks</td>
</tr>
<tr>
<td>p</td>
<td>significance level of r</td>
</tr>
</tbody>
</table>
Beyond structure

23 Bytes -> 32 Bytes

The cat sat on the mat.

El gato se sentó en la alfombra.
Content Similarity

23 Bytes -> 32 Bytes

The cat sat on the mat.

El gato se sentó en la alfombra.
Content Similarity

The cat sat on the mat.

El gato se sentó en la alfombra.

two-word-links  5
tsim = ------------  = ---
all links  8
Filtering with Features

Idea: Learn good/bad decision rule

Training data:
- Ask raters for content equivalence
- Positive examples easy

Challenges:
- Representative negative examples?
- Class skew
- Evaluation metric
Experiments on Internet Archive

- 120TB / 10B pages reported in 2003 paper!
- Setup similar to MapReduce (2004)
  - Map: For each URL generate language-independent URL
  - Shuffle: Group identical language-independent URLs together
  - Reduce: Match pairs, write parallel text
- Executed on 2 crawls of 8/12 TB
- Found ~8,300 EN-AR candidates
- After filtering: 1400 document pairs
Challenges

Translations on other sites
● siemens.com vs. siemens-systems.de
● News reported by different outlets

Machine Translation
● Too high scores look suspicious

Partial Translations
SEO (keywords in URLs)
North Korea executes defence minister for ‘dozing off’

13:05 05:06 CET

North Korea has executed its defence chief for treason according to South Korea’s spy agency. Leader Kim Jong-un has ordered the execution of 15 senior officials this year.

Hyon Yong-chol who was head of the country’s military is said to have shown his disloyalty to Kim by dozing off at a public event.

He was executed by firing squad – obliterated by an anti-aircraft gun – watched by hundreds of people, said South Korean media.

“Such a public and brutal method of execution would emphasise the cost of disloyalty.”

Analysts of the isolated county have observed that North Korean internal politics is very volatile at the moment.


Die sogenannte "Demokratische Volksrepublik Nordkorea" umfasst den Norden der koreanischen Halbinsel. Sie wird diktatorisch regiert und gilt als das gegenwärtig restriktivste politische System weltweit.
Some industry approaches
Focused crawling

Idea:
Collect domain specific corpus by browsing

Steps
1. Start with parallel page pair
2. Align structure
3. Follow the same link in both documents
4. Stop if new documents are not aligned
Dom tree alignment

A DOM tree alignment model for mining parallel data from the Web
Shi, Niu, Zhou, and Gao, Microsoft, ACL 2006

Key Idea:
Find alignment of nodes such that:
- If (parent) nodes A and B are aligned then
  - Children are also aligned (or deleted)
  - No reordering of children
DOM simplification

A Dom Tree        Our Document Tree
DOM Alignment Model

- Similar to synchronous tree substitution grammar (Hajic, 2004):
  - Support deletion / insertion / substitution
  - No reordering of children
  - Search for best alignment using DP

- Parameters:
  - Text-translation probabilities: $P(f|e)$
  - Tag-mapping probabilities, e.g. $P(<br>|<h1>)$
  - Node deletion probability
What Google does (or did in 2010)

For each non-English document:

1. Translate document into English using MT
2. Google it
3. Profit

New task:

Cross-lingual near duplicate detection
Near duplicate candidates

1. Find distinctive ngrams:
   a. rare, but not too rare (5-grams)
   b. used for matching only
2. Build inverted index:
   n-gram -> documents
N-gram* to document index

[cat sat on]  ->  {[doc_1, ES], [doc_3, DE], …}
[on the mat]  ->  {[doc_1, ES], [doc_2, ES], …}
[on the table]  ->  {[doc_3, DE]}

* Here: 3-grams, actually 5-grams
Matching using inverted index

[cat sat on]  ->  {[doc_1, ES], [doc_3, DE], ...}
[on the mat]  ->  {[doc_1, ES], [doc_2, ES], ...}
[on the table]  ->  {[doc_3, DE]}

For each n-gram:
  Generate all pairs where:
      document list short (<= 50)
      source language different

  {[doc_1, doc_3], ...}
Scoring using forward index

Forward index maps documents to n-grams

n = 2 for higher recall

For each document pair [d_1, d_2]:
- Collect scoring n-grams for both documents
- Build IDF-weighted vector
- Distance: cosine similarity
Conclusion

General pipeline:

- Find pairs
  - Within a single site / All over the Web
  - URL restrictions
  - IR methods

- Extract features
  - Structural similarity
  - Content similarity
  - Metadata

- Score pairs
Reading Material

Uszkoreit et al:
Large Scale Parallel Document Mining for Machine Translation, 2010

Resnik and Smith:
The Web as a Parallel Corpus, 2003

Shi et al:
A DOM tree alignment model for mining parallel data from the Web, 2006