Finding Chinese Abbreviation Pairs

Jeffrey Karres
January 23, 2006
Chinese Abbreviations

In Chinese, many words can be abbreviated by using several of its characters to stand for the entire word. For example:

<table>
<thead>
<tr>
<th>Chinese</th>
<th>Abbreviation</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>中國石油</td>
<td>中油</td>
<td>China Oil</td>
</tr>
<tr>
<td>台灣大學</td>
<td>台大</td>
<td>Taiwan University</td>
</tr>
<tr>
<td>高速鐵路</td>
<td>高鐵</td>
<td>high-speed rail</td>
</tr>
</tbody>
</table>
My goal:
Create a system that automatically finds abbreviation pairs
Is this easy to do?

Not yet.
Compositional?

- 台灣 = Taiwan
- 台 = terrace
- 灣 = bay
- 大學 = university
- 大 = big
- 學 = study
Compositional?

- 高速 = high-speed
- 高 = high
- 速 = speed
- 鐵路 = railroad
- 鐵 = iron
- 路 = road
Compositional?

- 台大 = terrace big?
- 高鐵 = high iron?
- Of course not!
- The point is that in order to understand these abbreviations, it is first necessary to know the full words of which they are abbreviations.
Pick out the abbreviations

• La plupart des miliciens de l’UTI ont été chassés...
• We are in MGH at UW.
• La ACLU argumenta que las medidas pisotean el poder...
• 以色列在約旦河西岸再次發生武裝衝突。

And while you’re at it, go ahead and tokenize...
If only the computer could do this for us...

• Should improve MT quality.

• Original Text: 因此便利迅捷的的電腦環境與高速的資訊轉換交流就是高鐵資訊系統所必備的要件

• Systran: Therefore facilitates the fast computer environment and the high speed information transformation exchange is the high-valence iron information system necessary important document

• Google: So the faster the computer environment and facilitate the rapid exchange of information is to change the elements that are necessary for the high-speed railway information system
If only...

- Information Retrieval
- Pretty much everything
What resources can we use?

- Monolingual Chinese text
- Chinese-English comparable corpora
- Chinese-English parallel corpora
Monolingual Chinese Text
Mining Atomic Chinese Abbreviation Pairs: A Probabilistic Model for Single Character Word Recovery

Jing-Shin Chang and Wei-Lun Teng
Chang and Teng

- Goal: find “atomic abbreviation pairs”
- What’s an atomic abbreviation pair?
  (中, 中國), (石油, 油), (網, 網路), (咖, 咖啡店),
  (高, 高雄), (雄, 高雄)
- The first of each pair is the abbreviation.
- The second is called the root.
C&T: Word Segmentation

- They combine the problems of abbreviation and segmentation.
- The idea is to look at a string of characters as the output of an HMM. (Remember the Korean tagger?)
- On the surface, we have characters.
- These characters actually come from hidden words.
C&T: Word Segmentation

• Hidden:
  • $c_1 = \text{中國}$
  • $c_2 = \text{石油}$

• Surface:
  • $w_l = \text{中油}$
C&T: Word Segmentation

- We want to find the best word sequence such that:

\[
\tilde{w}^* = \arg \max_{w_1^m : w_1^m \Rightarrow c_1^n} P(w_1^m | c_1^n)
\]

\[
= \arg \max_{w_1^m : w_1^m \Rightarrow c_1^n} P(c_1^n | w_1^m) \times P(w_1^m)
\]

\[
= \arg \max_{w_1^m : w_1^m \Rightarrow c_1^n} \prod_{i=1, m} P(\tilde{c}_i | w_i) \times P(w_i | w_{i-1})
\]
## C&T Word Segmentation

<table>
<thead>
<tr>
<th>characters:</th>
<th>台</th>
<th>大</th>
</tr>
</thead>
<tbody>
<tr>
<td>char. seq. cand. 1:</td>
<td>台灣</td>
<td>大學</td>
</tr>
<tr>
<td>char. seq. cand. 2:</td>
<td>台灣</td>
<td>大聯盟</td>
</tr>
</tbody>
</table>
C&T Word Segmentation

<table>
<thead>
<tr>
<th>characters:</th>
<th>台</th>
<th>大</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P(char. seq. cand. 1)</strong> = P(大學</td>
<td>台灣)</td>
<td>x P(台</td>
</tr>
<tr>
<td><strong>P(char. seq. cand. 2)</strong> = P(大聯盟</td>
<td>台灣)</td>
<td>x P(台</td>
</tr>
</tbody>
</table>
C&T’s ‘simplification’

\[
P(\tilde{c} \mid w) = P(c_1^m, \text{bit}, m \mid r_1^n, n)
\]

\[
\equiv P(c_1^m \mid r_1^n) \times P(\text{bit} \mid n) \times P(m \mid n)
\]

\[
\begin{cases} 
    c_1^m : \text{surface characters.} \\
    r_1^n : \text{root word characters.} \\
    m : \text{length of surface characters.} \\
    n : \text{length of root word characters.} \\
    \text{bit} : \text{bit pattern of abbreviation}
\end{cases}
\]
### C&T: bit pattern

<table>
<thead>
<tr>
<th>P(bit[n])</th>
<th>Score</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(10</td>
<td>2)</td>
<td>0.87</td>
</tr>
<tr>
<td>P(101</td>
<td>3)</td>
<td>0.44</td>
</tr>
<tr>
<td>P(1010</td>
<td>4)</td>
<td>0.56</td>
</tr>
<tr>
<td>P(10101</td>
<td>5)</td>
<td>0.66</td>
</tr>
<tr>
<td>P(101001</td>
<td>6)</td>
<td>0.51</td>
</tr>
<tr>
<td>P(1010001</td>
<td>7)</td>
<td>0.55</td>
</tr>
<tr>
<td>P(10101010</td>
<td>8)</td>
<td>0.21</td>
</tr>
</tbody>
</table>

**Table 1. High Frequency Abbreviation Patterns [by P(bit[n])]** (Chang and Lai, 2004)
C&T: Simplified Probabilities

<table>
<thead>
<tr>
<th>characters:</th>
<th>台</th>
<th>大</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(\text{char. seq. cand. 1})$ = $P(\text{大學</td>
<td>台灣})$</td>
<td>$\times P(台</td>
</tr>
<tr>
<td>$P(\text{char. seq. cand. 2})$ = $P(\text{大聯盟</td>
<td>台灣})$</td>
<td>$\times P(台</td>
</tr>
</tbody>
</table>

You can pick any combination of the p(char|word) probabilities (7 to choose from!)
The Single Character Recovery Model

- All Chinese abbreviations are either:
  - N characters -> 1 character, or
  - built up from abbreviations that are N-to-1

- So it would be a good idea to figure out all the basic N-to-1 abbreviations
How do we do this?

• Assume that each character is an abbreviated word. (A one-character word can be its own abbreviation.)

• Make a word lattice with the candidate root words (non-abbreviated forms) for each word.

• Find best path.

• Use this info to estimate parameters.

• Repeat until the best path no longer changes.
The Lattice
Another Lattice

Characters/abbreviations

Words

So here 台灣 doesn’t get counted as a word?
An Impossible Lattice?

Characters/abbreviations

Words

台灣
Questions

• So we’re not actually doing word segmentation after all?

• A (hidden) word can only emit a single character?

• So how exactly does the language model work?
Initial Abbreviation Pairs

• The initial pairs are generated by assuming that all words in a training corpus are potential ‘roots’ for any of the characters in that word.

• Furthermore, each single character is its own abbreviation.
Initial Probabilities

• If you can make sense of this, please explain it to me.

• How much effect on the end result will the initial probabilities have?
# Results

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Size</strong></td>
<td>13,086 KB</td>
<td>162</td>
</tr>
<tr>
<td><strong>Unique Abbrev. Patterns Found</strong></td>
<td>20,250</td>
<td>3,513</td>
</tr>
<tr>
<td><strong>Precision (picked sample of 100)</strong></td>
<td>62% at convergence</td>
<td>50%</td>
</tr>
</tbody>
</table>
Parallel Corpora
Using Parallel Corpora

NTU’s main campus is located in Taipei’s Daan district.

NTU’s medical school was established earlier than NTU itself.
NTU's main campus is located in Taipei’s Daan district.

NTU's medical school was established earlier than NTU itself.

台大 校總區 位於 台北市 大安 區

台灣 大學 醫學院 之 創立 比 台灣 大學 更早
Using Parallel Corpora

NTU's main campus is located in Taipei's Daan district.

NTU's medical school was established earlier than NTU itself.
Using Parallel Corpora

• Then find a way to decide if the pair in question really is an abbreviation pair.
Problems with this approach

- Sometimes the English will be abbreviated too! (But at least those are easier to find.)
- Limited amounts of parallel text
- Might be able to find abbreviations of named entities only.
Comparable Corpora
Using Comparable Corpora

- Find pieces of text that seem to deal with the same topic (based on vocabulary, time written, source)
- Find NE’s in the English and then try to find anything in the Chinese that might be a translation of that.
Weaknesses

• Given the procedure outline above, I will be finding only NE’s.

• Inherits all the weaknesses of NER in English.
Strengths

- Don’t have to wait until somebody makes a parallel corpus for you to use!
- The things that are found in news articles (which seem to be the main component in comparable text) are also the things we are likely to want to know about.
- We also get the English translation.
Evaluation

• Precision seems to be easier to calculate. However, numbers I get may not be comparable to those from the Chang and Teng paper.

• They are trying to find (char, word) pairs; I am looking for entire (2+ char) abbreviations

• Focus on NER
Evaluation

• Recall will be harder to find. Maybe look at unseen text and estimate what percentage of abbreviations have been identified as abbreviations.

• What counts as an abbreviation?