A Statistical Approach to Extract Chinese Chunk Candidates from Large Corpora

Zhang Le

www.nlplab.cn
Natural Language Processing Lab
Northeastern University, P.R. China
Organization

- Introduction
- Overall Procedure
- Fast Statistical Substring Reduction Algorithms
- Post Processing
- Evaluation
- Conclusion & Future Work
Organization

- Introduction
- Overall Procedure
- Fast Statistical Substring Reduction Algorithms
- Post Processing
- Evaluation
- Conclusion & Future Work
Introduction

Why traditional **bilingual translation unit** acquisition methods fail for Chinese?

- No large scale parsed corpus available
- Word segmentation problem of Chinese (and other oriental language)
- Sentence aligned bilingual corpus are hard to obtain
Our Approach

- Obtain chunk candidates from large monolingual corpora
- Extract bilingual translation unit from monolingual chunk candidates with the help of a small amount of annotated parallel corpus
- Using the acquired bilingual translation unit to promote translation result

I saw the heavy sea.

A Statistical Approach to Extract Chinese Chunk Candidates from Large Corpora – p.5/15
Our Approach

- Obtain chunk candidates from large monolingual corpora
- Extract bilingual translation unit from monolingual chunk candidates with the help of a small amount of annotated parallel corpus
- Using the acquired bilingual translation unit to promote translation result

I saw the heavy sea. <----> 我看见了波涛汹涌的大海.
Previous Work

The work of (Fung Pascale, 1994) showed: without the help of a machine-readable dictionary, the extracted trigrams and 4-grams from Chinese raw corpus contain only $31.3\%$ and $36.75\%$ valid phrases respectively.

A **Statistical Substring Reduction** procedure is required to filter out unnecessary n-gram sequences.
Statistical Substring Reduction

In order to rule out the majority “garbage strings” from the initial N-gram set, a *Statistical Substring Reduction* algorithm need to be employed to reduce some “garbage substrings” to their super strings.
Statistical Substring Reduction

In order to rule out the majority “garbage strings” from the initial N-gram set, a *Statistical Substring Reduction* algorithm need to be employed to reduce some “garbage substrings” to their super strings.

亚太经合组织 (Asia-Pacific Economic Cooperation) 10 times
亚太经合组 (Asia-Pacific Economic) 10 times
In order to rule out the majority “garbage strings” from the initial N-gram set, a Statistical Substring Reduction algorithm need to be employed to reduce some “garbage substrings” to their super strings.

亚太经合组织 (Asia-Pacific Economic Cooperation) 10 times
亚太经合组 (Asia-Pacific Economic) deleted 10 times
Statistical Substring Reduction

In order to rule out the majority “garbage strings” from the initial N-gram set, a *Statistical Substring Reduction* algorithm need to be employed to reduce some “garbage substrings” to their super strings.

Asia-Pacific Economic Cooperation (Asia-Pacific Economic) 10 times

Since the latter is the substring of the former with the same frequency. This procedure is called *Statistical Substring Reduction*, which reduces some “garbage substrings” to their super strings using frequency information.
A Simple SSR Algorithm

Traditional Statistical Substring Reduction algorithm (Han et al., 2001) is an $O(n^2)$ algorithm and unable to handle large corpora.

1: for $i = 1$ to $n$ do
2:  for $j = 1$ to $n$ do
3:    if $X_i \propto X_j$ and $f_i - f_j < k$ then
4:      $M_i = 1$
5:    end if
6:  end for
7: end for
Two Fast SSR Algorithms $O(n)$

To address the time problem in traditional SSR algorithm, we proposed two new Fast Statistical Substring Reduction (FSSR) algorithms, both have an $O(n)$ time complexity under ideal condition. (LÜ 2003) gives a mathematical proof on the equality of four SSR algorithms.
Post Processing

After performing SSR operation on extracted N-gram set, a post processing procedure is carried out to do some further filtering.

- Mutual Information Filtering
- Stopword List
- Language Specific Treatment (word length, etc.)

Post processing method is simple and effective for this task.
## Performance of FSSRs

We perform three SSR algorithms on three corpora of different sizes (2 - 20-gram):

<table>
<thead>
<tr>
<th>Label</th>
<th>Time (Including I/O)</th>
<th>Algo 1</th>
<th>Algo 2</th>
<th>Algo 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>corpus1 (3.5MB)</td>
<td>17 min 20 sec</td>
<td>3.3 sec</td>
<td>4.4 sec</td>
<td></td>
</tr>
<tr>
<td>corpus2 (50MB)</td>
<td>27 hours</td>
<td>48.8 sec</td>
<td>54.6 sec</td>
<td></td>
</tr>
<tr>
<td>corpus3 (1GB)</td>
<td>N/A</td>
<td>8 min 23 sec</td>
<td>7 min 25 sec</td>
<td></td>
</tr>
</tbody>
</table>

Even on small corpus like corpus1, the two FSSRs are 200 - 300 times faster than traditional SSR algorithm.
Performance of FSSRs

We perform three SSR algorithms on three corpora of different sizes (2 - 20-gram):

<table>
<thead>
<tr>
<th>Label</th>
<th>Time (Including I/O)</th>
<th>Algo 1</th>
<th>Algo 2</th>
<th>Algo 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>corpus1 (3.5MB)</td>
<td>17 min 20 sec</td>
<td>3.3 sec</td>
<td>4.4 sec</td>
<td></td>
</tr>
<tr>
<td>corpus2 (50MB)</td>
<td>27 hours</td>
<td>48.8 sec</td>
<td>54.6 sec</td>
<td></td>
</tr>
<tr>
<td>corpus3 (1GB)</td>
<td>N/A</td>
<td>8 min 23 sec</td>
<td>7 min 25 sec</td>
<td></td>
</tr>
</tbody>
</table>

Even on small corpus like corpus1, the two FSSRs are **200 - 300** times faster than traditional SSR algorithm.
Extraction Result

Manually checking 1000 candidate n-gram sequences randomly: **86.3%** are meaningful chunk candidates. Some results from PeopleDaily 2000 corpus:

<table>
<thead>
<tr>
<th>Meaningful Chunks</th>
<th>Nonsensical Chunks</th>
</tr>
</thead>
<tbody>
<tr>
<td>被偷 (be stolen)</td>
<td>丽画</td>
</tr>
<tr>
<td>明确地表示 (to express explicitly)</td>
<td>院所属</td>
</tr>
<tr>
<td>可口可乐公司 (the Coca-Cola company)</td>
<td>处寻找</td>
</tr>
<tr>
<td>发展民族教育 (developing national education system)</td>
<td>著名女</td>
</tr>
<tr>
<td>语重心长地说 (to tell with great patience)</td>
<td>明确保</td>
</tr>
<tr>
<td>瓦斯爆炸事故 (gas explosion accident)</td>
<td>成社会主义</td>
</tr>
<tr>
<td>义务植树活动 (tree-planting action by volunteers)</td>
<td>量逐年增</td>
</tr>
<tr>
<td>遇到许多困难 (come across many difficulties)</td>
<td>通过了专家</td>
</tr>
<tr>
<td>增进了相互了解和友谊 (to improve the friendship and mutual understanding)</td>
<td>推动两岸人员往来和各</td>
</tr>
</tbody>
</table>
Conclusion

Highlights of our method:

- Purely statistical method (Language in-depend, no human intervention)
- Efficient & Effective (two FSSRs)
- Encouraging result (35% → 85%)
Conclusion

Highlights of our method:

- Purely statistical method (Language in-depend, no human intervention)
- Efficient & Effective (two FSSRs)
- Encouraging result (35% → 85%)

Drawbacks:

- Not 100% accurate, some meaningful chunk candidates are discarded
- Post processing is too simple
- Not linguistic aware
Future Work

Some perspective:

- Integration of Statistical Language Model (SLMs)
- Resort to shallow parsing technology (POS, NP Chunk, etc.)
- Proper name identification
This is the End, Thank you!