Iterative Mining Translations from the Web

Fang Li, Shuangqing Yuan, and Huanye Sheng

Reporter: 江欣倩

Professor: 陳嘉平

Outline

- Introduce
- former research on mining translation pairs on the Internet
- Iterative mining approach

Outline

- **Introduce**
- former research on mining translation pairs on the Internet
- Iterative mining approach

Introduction

- The number of bilingual or multilingual web pages is increasing faster than people’s expectation.
  - Parallel web pages
  - Comparable web pages
- Approach of mining translations on line
  - Alignment methods (STRAND)
  - Hyperlink methods
  - Other methods

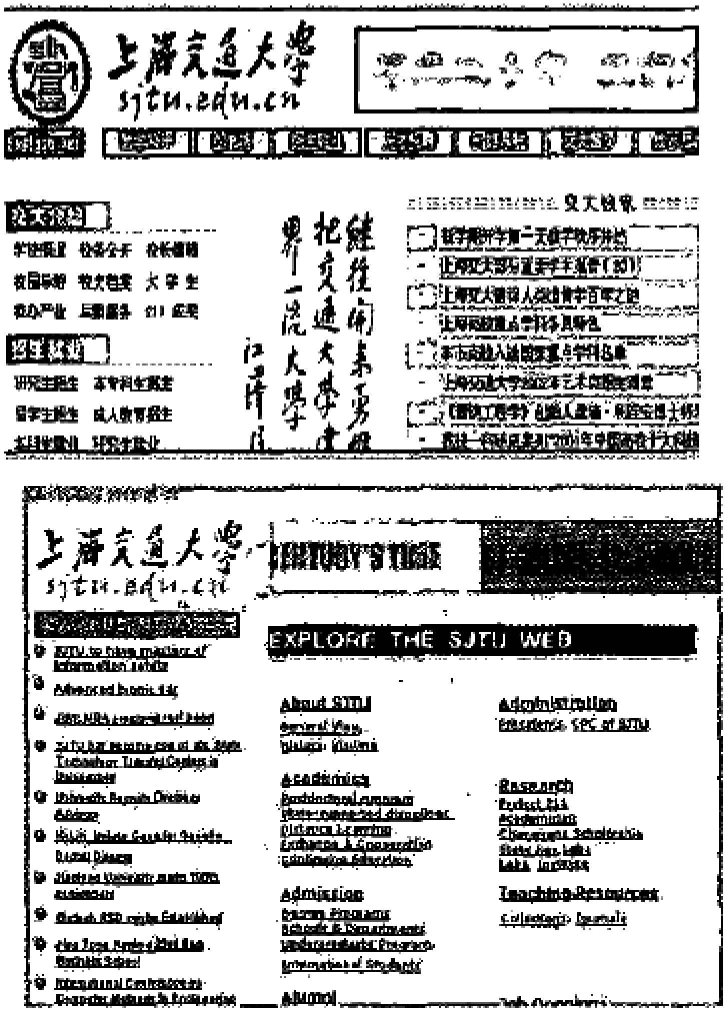
Outline

- Introduce
- **former research on mining translation pairs on the Internet**
- Iterative mining approach

Web page components

- Four components
  - Title of the web page
  - Text of the web page
  - Markups (tags)
  - Hyperlinks
    - <a href=“http://www.lib.sjtu.edu.cn”>圖書館
    - <a href=“http://www.lib.sjtu.edu.ch/english/”>Libraries

Fig.1. Unparallel web pages



上海交通大學
sjtu.edu.cn

上海交通大學
sjtu.edu.cn

EXPLORE THE SJTU WEB

About SJTU

Administration

Academics

Research

Admission

Teaching Resources

Alumni

Algorithm

1. Find the anchor point of two hyperlink vectors and DIS
2. Calculate Length penalty
3. Calculate the similarity of each term
4. Calculate the final result

Hyperlink vector

- Separate hyperlink by “/”
  - $H_e(T_1,T_2,\ldots,T_n)$
    - <a href=“http://www.lib.sjtu.edu.cn/english/”>
    - H_e(“http:”,”www.lib.sjtu.edu.cn”,”english”)
  - $H_c(M_1,M_2,\ldots,M_n)$
    - <a href=“http://www.lib.sjtu.edu.cn”>
    - H_c(“http:”,”www.lib.sjtu.edu.cn”)

Find out the anchor point for two hyperlinks

- The form of a hyperlink
    - Complete path
        - http://www.chian-cts.com/service/tele.htm
    - Relative path
        - service/tele.htm
    - Starting from its parent path
        - ../service/tele.htm

- Example
    - <a href="sitemap/index.html">網站地圖
      <a href="../eng/sitemap/index.html">sitemap
    - The anchor point is "sitemap"
        - H_c("sitemap","index.html")
          H_c("eng","sitemap","index.html")
            - The displacement (DIS) in a example is 1.
        - $H_{cc}(T_1,T_2,\ldots,T_n)$ and $H_{ee}(M_1,M_2,\ldots,M_m)$
            - H_c("sitemap","index.html")
              H_c("sitemap","index.html")

Length Penalty and Dice coefficient

- Length Penalty

$$Length\_Penalty = \frac{2 \times \min(n, m)}{m + n}$$

- The similarity of each term
  - Dice coefficient

$$Dice(T_i, M_i) = 2 \times \frac{\left|T_i \cap M_i\right|}{\left|T_i\right| + \left|M_i\right|} = sim_i$$

    - $\left|T_i \cap M_i\right|$: the length of common sub-string of T_i and M_i
    - $|M_i|$, $|T_i|$: the lengths of M_i and T_i individually

The similarity of H_{ee} and H_{cc}

- $sim(H_{ee}, H_{cc}) = Length\_Penalty \times \frac{1}{DIS} \times \sqrt{\frac{\sum_{i=1}^{\min(n,m)} sim_i^2}{\min(n,m)}}$

Evaluation

- Hyperlinks from about 100 web sites with Chinese and English version
  - 2567 Chinese hyperlinks
  - 771 English hyperlinks
  -

Similarity	Total items	Correct items	Precision
1	50	45	90%
0.9-1	18	18	100%
0.8-0.9	22	20	90%
0.7-0.8	15	7	47%
0.5-0.7	13	4	38%
<0.5	58	2	5%

Outline

- Introduce
- former research on mining translation pairs on the Internet
- **Iterative mining approach**

Definition and Principle

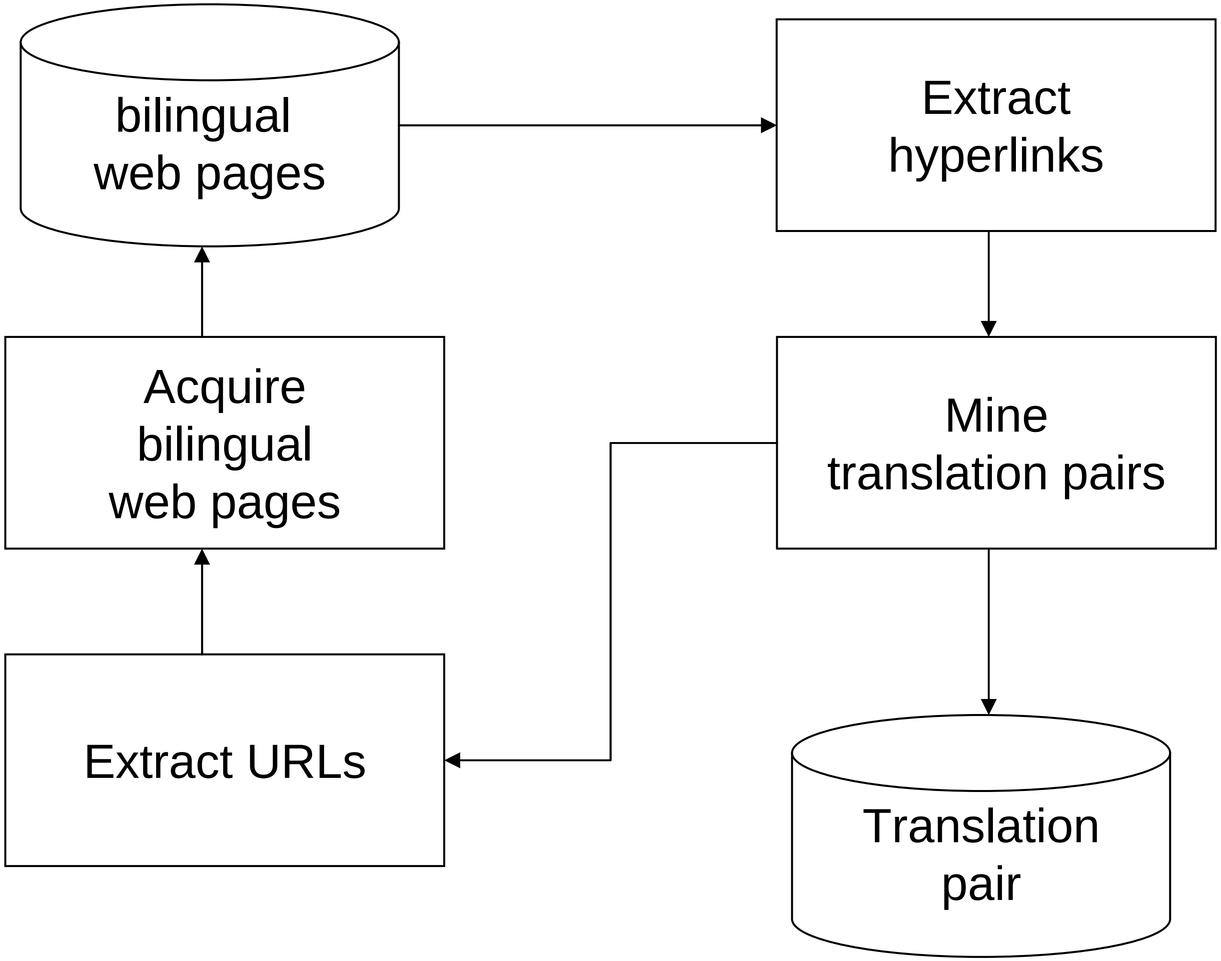
- Definition
  - a hyperlink triple
    <Content, SrcURL, HrefValue>
    - example
      - <“上海交通大學計算機系”, http://www.sjtu.edu.cn, “http://cs.sjtu.edu.cn”>
- Principle
  - Link_1<Content_1, SrcURL_1, HrefValue_1>
    Link_2<Content_2, SrcURL_2, HrefValue_2>
    1. the more similar of HrefValue_1 and HrefVaule_2, the more likely that Content_1 and Content_2 become translation
    2. the higher the weight of translation pair <Content_1, Content_2>, the more likely that HrefValue_1 and HrefValue_2 become a pair of bilingual page

Definition and Principle

- Definition
  - a hrefvalue triple
    <protocol, hostname, pathList>
    - pathList (p_1\\$p_2$\\...\\$p_n$)
      - <p_1,p_2,…,p_n>
  - prefix or postfix in path
    - example
      - <a href=“e_other/rmbcredit.jsp”>
    - hyperlinks from 333 pairs Chinese-English web pages
      -

	Number of hyperlink	With prefix or postfix	rate
Chinese Web page	15,695	3,896	24.8%
English Web page	12,211	3,963	32.5%

The system architecture



Algorithm of Extracting Translation Pairs

- ExtractTP(Link1<Content_1,SrcURL_1,HrefValue_1>, Link2<Content_2,SrcURL_2,HrefValue_2>)

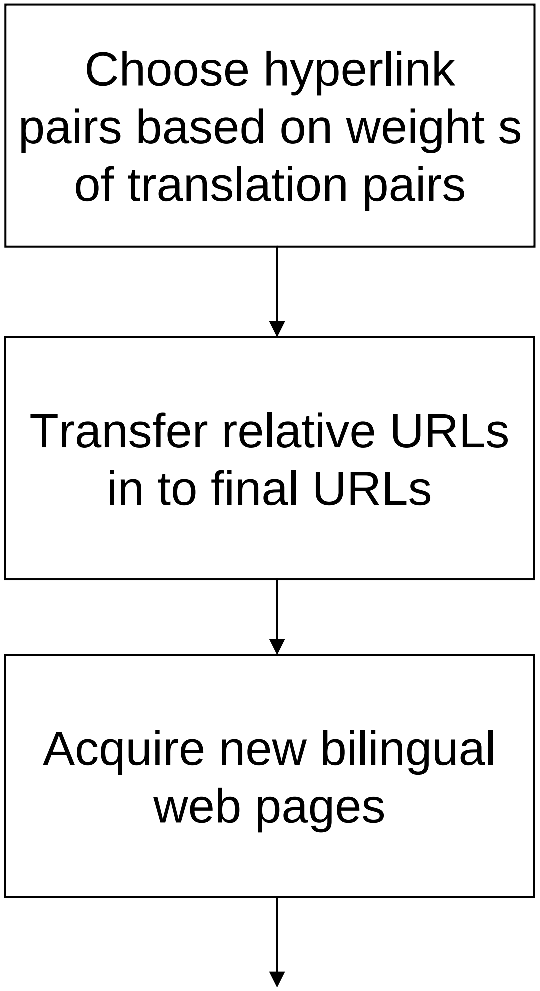
  1. form two sets of links $\textit{Link-set}_1$ and $\textit{Link-set}_2$ from web page $\textit{SrcURL}_1$ and $\textit{SrcURL}_2$
  2. compare Link_i in $\textit{Link-Set}_1$ with Link_j in $\textit{Link-Set}_2$ (the similarity score)
  3. if score is greater than the threshold (0.5), Content_1 and Content_2 are regarded as translations of each other

The Similarity

- N: the number of equal parts
  M: the number of not equal
  S: N/(N+M)

```
if $hostname_i$ != $hostname_j$ then S = 0;
else if
$pathList_i$ or $pathList_j$ has prefix or postfix, then
Delete its prefix or postfix
else if
$p_i$ (in $pathList_i$<$p_1$,$p_2$,…,$p_n$>) == $q_i$ (in $pathList_j$<$q_1$,$q_2$,…,$q_n$>)
then N++, else M++
```

Bilingual Web pages acquisition



Choose hyperlink pairs based on weight s of translation pairs

Transfer relative URLs in to final URLs

Acquire new bilingual web pages

- If the weight (similarity) of any translation is greater than or equal to the threshold (1), their hyperlinks are regarded as potential bilingual web pages.

Results and analysis

- Initial 333 bilingual (Chinese-English) web pages
- threshold
  - for translation pairs: 0.5
  - for new bilingual web pages: 1
- Results

Iteration	New bilingual web page pairs	Bilingual translation pairs	Precision of translation pairs
1	453	775	89.1%
6	898	1261	90.2%

Hyperlinks	Precision
Without deleting prefix /postfix	74.5%
Deleting prefix/postfix	90.2%

Conclusion

- mine translation from bilingual web pages
- This paper introduces two new features
  - get more translation pairs to improve hyperlink-based method by iterative process
  - prefix/postfix filter processing