An Iterative Link-based Method for Parallel Web Page Mining

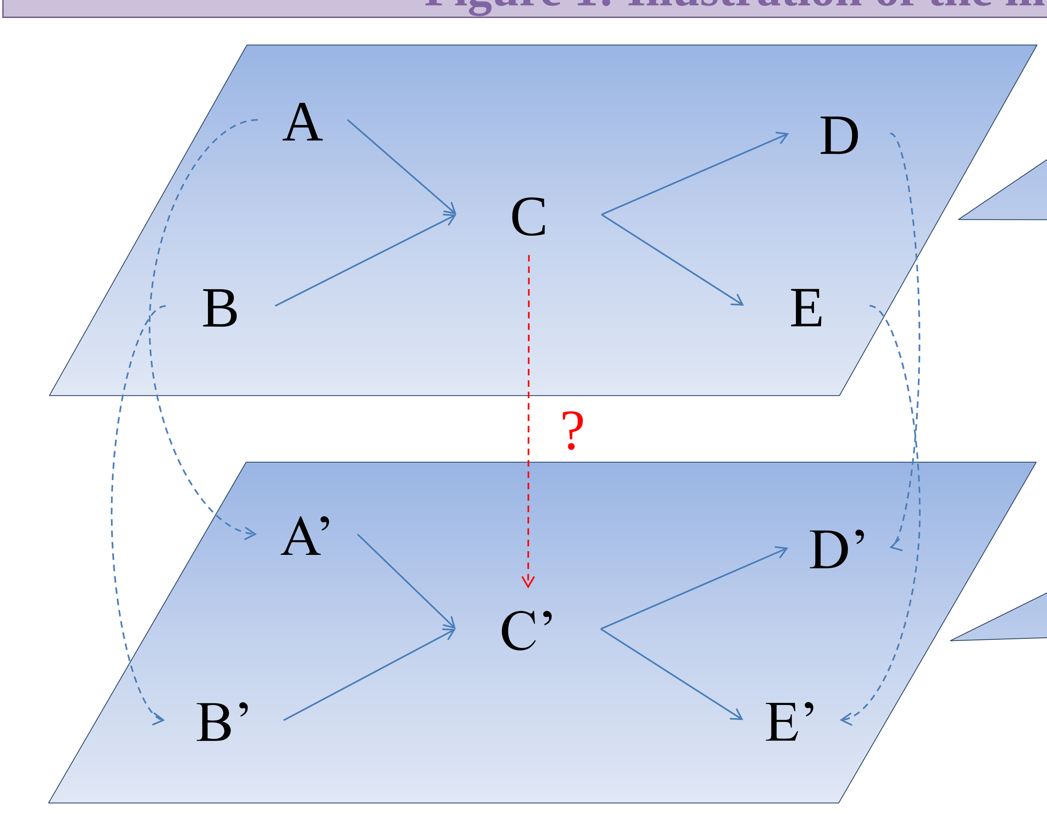
Le Liu[1], Yu Hong[1], Jun Lu[2], Jun Lang[2], Heng Ji[3], Jianmin Yao[1]
[1]*Natural Language Processing Lab, Soochow University, Suzhou, 215006, China*
[2]*Institute for Infocomm Research, Singapore, 138632*
[3]*Computer Science Deparment, Resselaer Polytechnic Institue, Troy, NY 12180, USA*
{leliuchn, tianxianer, lujun59, billlangjun}@gmail.com, jih@rpi.edu, jyao@suda.edu.cn

Introduction

- In this paper, we address the task of *parellel web page mining* by first using *hyperlink information* of web pages within bilingual website.
- We propose an *iterative link-based approach* which combines both *internal and external translation similarity* of web pages to identify parallel web page.

Motivation

Figure 1: Illustration of the link-based method



A, B, C, D and *E* are web pages in the same language; while *A', B', C', D'* and *E'* are web pages in another language. (e.g. Page *A, B, C, D, E* are pages in English while Page *A', B', C', D', E'* are in Chinese)

The solid arrows are the links between these pages within bilingual website. The dashed arrows indicate that page pair has high translation similarity.

We hypothesize that page pairs <*C,C'*> might be parallel web pages if page *C*'s neighbors {*A, B, D, E*} have a higher translation similarity with page *C'* 's neighbors {*A', B', D', E'*} respectively.

Algorithm Flow

Algorithm 1: Estimating the external translation similarity

Input: $PG(e), PG(c)$
Output: $S_{ext}^{i}(e,c)$
Procedure:
$sum \leftarrow 0$
$e\_set \leftarrow PG(e)$
$c\_set \leftarrow PG(c)$
While e_set and c_set are both not empty:
$<x,y> \leftarrow arg\,max_{x\in e\_set,\, y\in c\_set}\,(ETS^{i-1}(x,y))$
$sum \leftarrow sum + ETS^{i-1}(x,y)$
Remove x from e_set
Remove y from c_set
$S_{ext}^{i}(e,c) = Sim^{i}(p(e),p(c)) = 2 \cdot sum\,/\,(|PG(e)| + |PG(c)|)$

Here, $S_{ext}^{i}(e,c)$ is the external translation similarity of page e and c after the i-th iteration, and the same is for $ETS^{i-1}(e,c)$ and $Sim^{i}(p(e),p(c))$. $|PG(x)|$ is the number of x's neighbors.

Algorithm 2 Estimating the enhanced translation similarity

Input: P_e, P_c, *(the English and Chinese page set)*
Output: $ETS(e,c), e \in P_e, c \in P_c$
Initialization:
Set $ETS(e,c)$ random value or small value
Procedure:
LOOP:
For each e in P_e:
For each c in P_c:
$ETS^{i}(e,c) = \alpha \cdot S_{ext}^{i}(e,c) + (1-\alpha) \cdot S_{in}(e,c)$
Parameters normalization
UNTIL $ETS(e,c)$ is stable

The Baseline system only adopts internal information of web pages for identifying parallel web pages. Figure 2 shows that when β is set to 0.6, the baseline system achieves the best performance. Thus, we always set β to 0.6. Figure 4 shows that when the parameter α is set to 0.6, our method achieves the best performance and obtains significant improvement (6.2% F-score) over the baseline system. The experimental results show that the external information of web pages is an effective feature to mine parallel web pages.

Algorithm 3 Finding parallel page pairs

Input: $P_e, P_c, ETS(x,y), x \in P_e, y \in P_c, MAX\_P$ (or MIN_SIM)
Output: Parallel Page Pairs List : PPL
Procedure:
LOOP:
$<x,y> = arg\,max_{x\in P_e, y\in P_c}\,(ETS(x,y))$
Add $<x,y>$ to PPL
Remove x from P_e
Remove y from P_c
UNTIL size of $PPL > MAX\_P$ (or $ETS(x,y) < MIN\_SIM$)

The input MAX_P is an integer threshold which means that only top MAX_P page pairs will be extracted in a certain website.

Figure 3 shows that the performance of our method achieves the maximal values and converges after the third iteration. In addition, Figure 3 indicates that our method is robust for different websites. Thus, The iteration number is set to 3 in the following experiments.

Model Definition

Enhanced Translation Similarity

$$ETS(e,c) = (1-\alpha) \cdot S_{in}(e,c) + \alpha \cdot S_{ext}(e,c), \alpha \in [0,1]$$

Internal Translation Similarity

$$S_{in}(e,c) = \beta \cdot S_{cb}(e,c) + (1-\beta) \cdot S_{struct}(e,c), \beta \in [0,1]$$

External Translation Similarity

$$S_{ext}(e,c) = Sim(PG(e), PG(c))$$

- $S_{in}(e,c)$ is the internal translation similarity of two pages: e and c. $S_{ext}(e,c)$ is the external translation similarity of pages e and c. $ETS(e,c)$ is Enhanced Translation Similarity of two pages, which combines internal with external translation similarity to identify parallel web pages.
- $S_{in}(e,c)$ is the internal translation similarity of pages e and c which combines content-based similarity $S_{cb}(e,c)$ and structural similarity $S_{struct}(e,c)$ with linear weight. Here, $S_{cb}(e,c)$ is the percentage of translation word pairs in two pages with a small bilingual lexicon. $S_{struct}(e,c)$ is the longest common sequences of two HTML tag sequences in page e and c.
- PG(x),which is a set of pages, is the neighbors of page x. $S_{ext}(e,c)$ indicates the external translation similarity of pages e and c. Here, it is the similarity of two page set PG(e) and PG(c), which relies on the similarity of the elements in the page set. Thus, $Sim(PG(e),PG(c))$ depends on $ETS(e_i,c_j)$ (e_i, c_j belongs to $PG(e)$, $PG(c)$, respectively) and $ETS(e,c)$. $ETS(e,c)$ depends on $S_{in}(e,c)$ and $S_{ext}(e,c)$. It is an iterative process.

Experimentation

- We conduct our experiments on six bilingual websites which are selected from HK government websites.The test data is randomly extracted from these websites and annoted by URL-based pattern rules and human annotator.
- All the web pages are retrieved by using a web site download tool: HTTrack.
- We adopt Precision, Recall and F-score to evaluate our method.

Table 1: Number of pages and bilingual page pairs of each websites

Site ID	En/Ch pages	Total pairs	No pattern pairs	URL
S1	1101/1098	1092	20	www.gov.hk
S2	501/497	487	7	www.customs.gov.hk
S3	995/775	768	12	www.sbc.edu.sg
S4	4085/3838	3648	4	www.swd.gov.hk
S5	660/637	637	0	www.landsd.gov.hk
S6	4733/4626	4615	8	www.td.gov.hk
total	12075/11471	11684	51	

Figure 2: Performances of baseline system with different β value

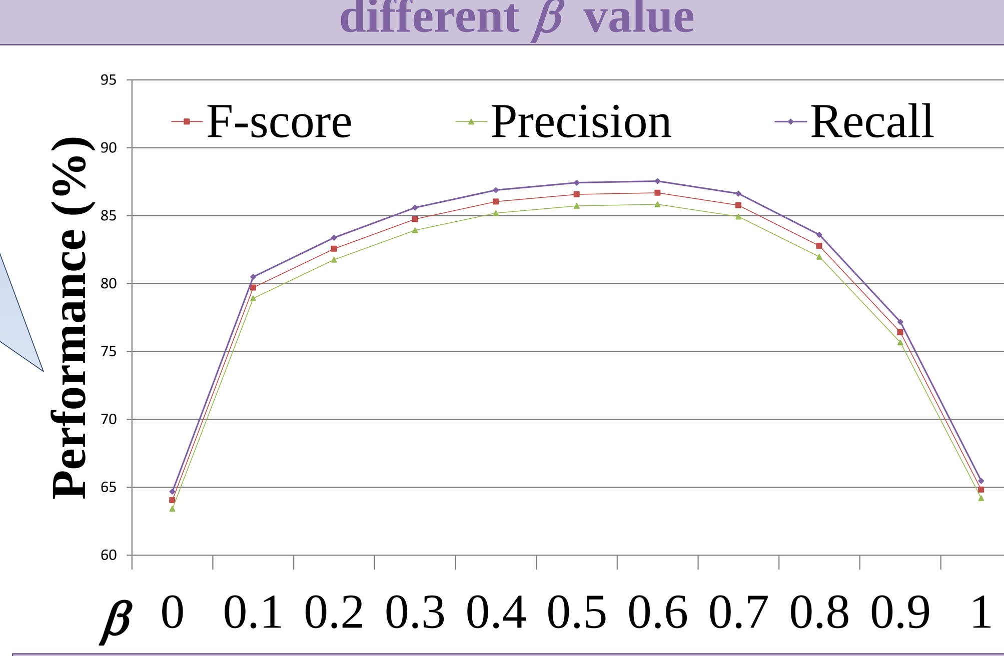


Figure 4: The F-scores of our method with different the value of α

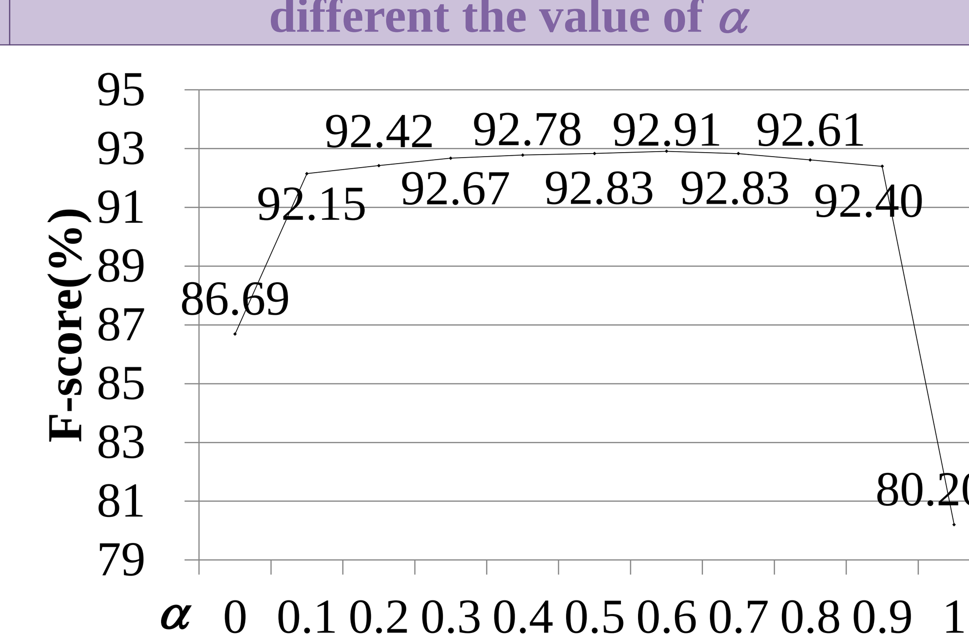


Figure 3: Experiment results of our method on each website