Available online at www.sciencedirect.com

SCIENCE DIRECT®

Information Processing and Management 42 (2006) 1294–1315

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Automatic extraction of bilingual word pairs using inductive chain learning in various languages ☆

Hiroshi Echizen-ya [a,*], Kenji Araki [b], Yoshio Momouchi [a]

[a] *Department of Electronics and Information Engineering, Hokkai-Gakuen University, South-26 West-11, Chuo-ku, Sapporo 064-0926, Japan*
[b] *Graduate School of Information Science and Technology, Hokkaido University, Kita-14, Nishi-9, Kita-ku, Sapporo 060-0814, Japan*

Received 31 July 2005; received in revised form 23 November 2005; accepted 30 November 2005
Available online 23 January 2006

Abstract

In this paper, we propose a new learning method for extracting bilingual word pairs from parallel corpora in various languages. In cross-language information retrieval, the system must deal with various languages. Therefore, automatic extraction of bilingual word pairs from parallel corpora with various languages is important. However, previous works based on statistical methods are insufficient because of the sparse data problem. Our learning method automatically acquires rules, which are effective to solve the sparse data problem, only from parallel corpora without any prior preparation of a bilingual resource (e.g., a bilingual dictionary, a machine translation system). We call this learning method Inductive Chain Learning (ICL). Moreover, the system using ICL can extract bilingual word pairs even from bilingual sentence pairs for which the grammatical structures of the source language differ from the grammatical structures of the target language because the acquired rules have the information to cope with the different word orders of source language and target language in local parts of bilingual sentence pairs. Evaluation experiments demonstrated that the recalls of systems based on several statistical approaches were improved through the use of ICL.

Keywords: Learning method; Bilingual word pairs; Various languages; Sparse data problem; Parallel corpora; Statistical approach

1. Introduction

1.1. Sparse data problem in parallel corpora

In the field of Cross-Language Information Retrieval (CLIR) (Chen & Gey, 2004; Fujii & Ishikawa, 2001; Kishida et al., 2004; Xu & Weischedel, 2003), bilingual word pairs—the pairs of Source Language (SL) words and Target Language (TL) words—are essential. However, manual extraction by humans of bilingual word

☆ Note: An earlier version of this paper was presented at the NLDB2005 Conference.
[*] Corresponding author. Tel.: +81 11 841 1161x7863; fax: +81 11 551 2951.
E-mail addresses: echi@eli.hokkai-s-u.ac.jp (H. Echizen-ya), araki@media.eng.hokudai.ac.jp (K. Araki), momouchi@eli.hokkai-s-u.ac.jp (Y. Momouchi).

0306-4573/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2005.11.004

Table 1
A contingency matrix

	W_T	$\neg W_T$	
W_S	a	b	
$\neg W_S$	c	d	
			n

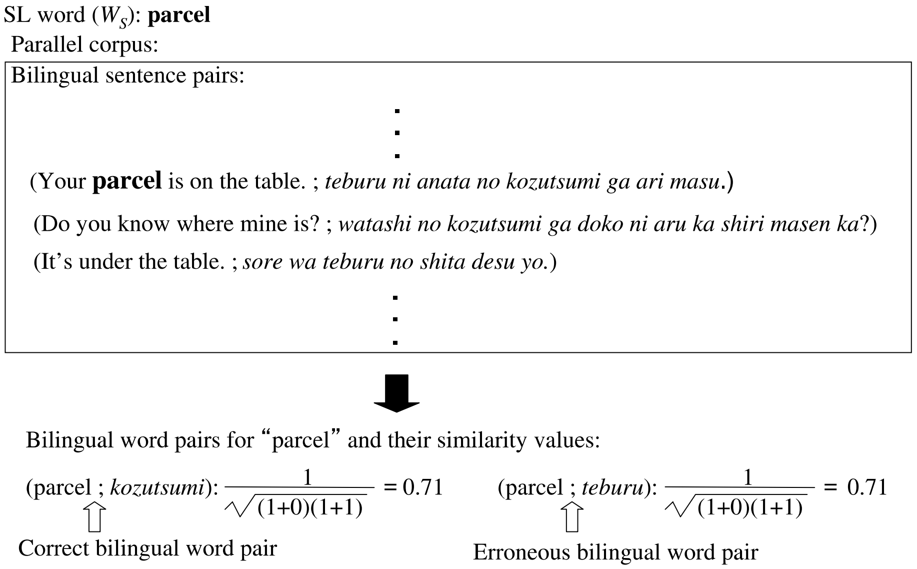
SL word (W_S): **parcel**
Parallel corpus:
Bilingual sentence pairs:
(Your **parcel** is on the table. ; *teburu ni anata no kozutsumi ga ari masu.*)
(Do you know where mine is? ; *watashi no kozutsumi ga doko ni aru ka shiri masen ka*?)
(It's under the table. ; *sore wa teburu no shita desu yo.*)
Bilingual word pairs for "parcel" and their similarity values:
(parcel ; *kozutsumi*): $\frac{1}{\sqrt{(1+0)(1+1)}} = 0.71$
(parcel ; *teburu*): $\frac{1}{\sqrt{(1+0)(1+1)}} = 0.71$
Correct bilingual word pair
Erroneous bilingual word pair

Fig. 1. An example of sparse data problem by the system based on cosine.

pairs in various languages is costly. For that reason, automatic extraction of bilingual word pairs from parallel corpora in various languages is important to make the method feasible. Statistical approaches are effective to extract bilingual word pairs from parallel corpora with various languages (Kay & Röscheisen, 1993; Manning & Schütze, 1999; Melamed, 2001; Sadat, Yoshikawa, & Uemura, 2003; Smadja, McKeown, & Hatzivassiloglou, 1996; Veronis, 2000) because they are language independent. However, they are insufficient because of sparse data problem. For example, the system uses cosine (Manning & Schütze, 1999) to extract bilingual word pairs from a parallel corpus. The cosine is the representative similarity measure in the statistical approaches. The cosine is defined as

$$\text{cosine}(W_S, W_T) = \frac{a}{\sqrt{(a+b)(a+c)}} \tag{1}$$

Table 1 shows the parameters used in function (1): W_S is an SL word and W_T is a TL word in a parallel corpus. The number of pieces in which both W_S and W_T were found in each bilingual sentence pair is represented as 'a'; 'b' is the number of pieces in which only W_S was found in each bilingual sentence pair; and 'c' is the number of pieces in which only W_T was found in each bilingual sentence pair. In addition, 'n' represents the total number of words in a parallel corpus; 'd' denotes the values of '$n - (a + b + c)$'.

Fig. 1 shows an example of the sparse data problem by the system based on cosine.

In Fig. 1, the system based on cosine cannot extract only bilingual word pair (parcel;*kozutsumi*[1]) because the similarity value between "parcel" and "*kozutsumi*" becomes 0.71, and the similarity value between "parcel" and "*teburu*" also becomes 0.71 by the cosine function (1). This problem becomes very serious when the

[1] Italics indicate Japanese pronunciation. Space (i.e. ' ') in Japanese sentences are inserted after each morpheme because Japanese is an agglutinative language. This process is automatically performed using the Japanese morphological analysis system "ChaSen" (Matsumoto et al., 2000). Grammatical structure of Japanese is SOV.