Expanding Indonesian-Japanese Small Translation Dictionary Using a Pivot Language

Masatoshi Tsuchiya[†] **Ayu Purwarianti**[‡] **Toshiyuki Wakita**[‡] **Seiichi Nakagawa**[‡]
[†]Information and Media Center / [‡]Department of Information and Computer Sciences,
Toyohashi University of Technology
tsuchiya@imc.tut.ac.jp, {wakita,ayu,nakagawa}@slp.ics.tut.ac.jp

Abstract

We propose a novel method to expand a small existing translation dictionary to a large translation dictionary using a pivot language. Our method depends on the assumption that it is possible to find a pivot language for a given language pair on condition that there are both a large translation dictionary from the source language to the pivot language, and a large translation dictionary from the pivot language to the destination language. Experiments that expands the Indonesian-Japanese dictionary using the English language as a pivot language shows that the proposed method can improve performance of a real CLIR system.

1 Introduction

Rich cross lingual resources including large translation dictionaries are necessary in order to realize working cross-lingual NLP applications. However, it is infeasible to build such resources for all language pairs, because there are many languages in the world. Actually, while rich resources are available for several popular language pairs like the English language and the Japanese language, poor resources are only available for rest unfamiliar language pairs.

In order to resolve this situation, automatic construction of translation dictionary is effective, but it is quite difficult as widely known. We, therefore, concentrate on the task of expanding a small existing translation dictionary instead of it. Let us consider three dictionaries: a small *seed dictionary* which consists of headwords in the source language and their translations in the destination language, a large *source-pivot dictionary* which consists of headwords in the source language and their translations in the pivot language, and a large *pivot-destination dictionary* which consists of headwords in the pivot language and their translations in the destination language. When these three dictionaries are given, expanding the seed dictionary is to translate words in the source language that meets two conditions: (1) they are not contained in the seed dictionary, and (2) they can be translated to the destination language transitively referring both the source-pivot dictionary and the pivot-destination dictionary.

Obviously, this task depends on two assumptions: (a) the existence of the small seed dictionary, and (b) the existence of the pivot language which meets the condition that there are both a large source-pivot dictionary and a large pivot-destination dictionary. Because of the first assumption, it is true that this task cannot be applied to a brand-new language pair. However, the number of such brand-new language pairs are decreasing while machine-readable language resources are increasing. Moreover, The second assumption is valid for many language pairs, when supposing the English language as a pivot. From these point of view, we think that the expansion task is more promising, although it depends more assumptions than the construction task.

There are two different points among the expansion task and the construction task. Previous researches of the construction task can be classified into two groups. The first group consists of researches to construct a new translation dictionary for a fresh language pair from existing translation dictionaries or other language resources (Tanaka and Umemura, 1994). In the first group, information of the seed dictionary are not counted in them unlike the expansion task, because it is assumed that there is no seed dictionary for such fresh language pairs. The second group consists of researches to translate

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 197–200,
Prague, June 2007.

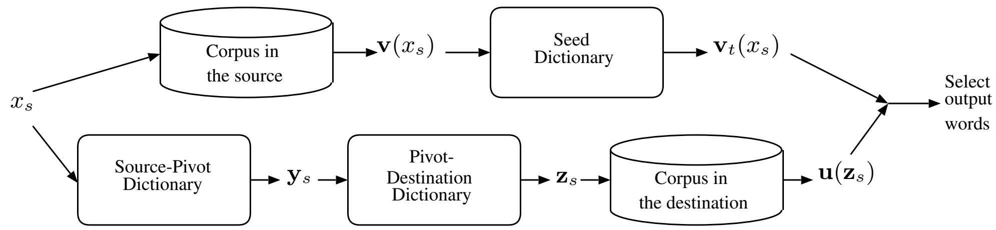

Figure 1: Translation Procedure

novel words using both a large existing translation dictionary and other linguistic resources like huge parallel corpora (Tonoike et al., 2005). Because almost of novel words are nouns, these researches focus into the task of translating nouns. In the expansion task, however, it is necessary to translate verbs and adjectives as well as nouns, because a seed dictionary will be so small that only basic words will be contained in it if the target language pair is unfamiliar. We will discuss about this topic in Section 3.2.

The remainder of this paper is organised as follows: Section 2 describes the method to expand a small seed dictionary. The experiments presented in Section 3 shows that the proposed method can improve performance of a real CLIR system. This paper ends with concluding remarks in Section 4.

2 Method of Expanding Seed Dictionary

The proposed method roughly consists of two steps shown in Figure 1. The first step is to generate a co-occurrence vector on the destination language corresponding to an input word, using both the seed dictionary and a monolingual corpus in the source language. The second step is to list translation candidates up, referring both the source-pivot dictionary and the pivot-destination dictionary, and to calculate their co-occurrence vectors based on a monolingual corpus in the destination.

The seed dictionary is used to convert a co-occurrence vector in the source language into a vector in the destination language. In this paper, $f(w_i, w_j)$ represents a co-occurrence frequency of a word w_i and a word w_j for all languages. A co-occurrence vector $\mathbf{v}(x_s)$ of a word x_s in the source is:

$$\mathbf{v}(x_s) = (f(x_s, x_1), \ldots, f(x_s, x_n)), \qquad (1)$$

where $x_i (i = 1, 2, \ldots, n)$ is a headword of the seed dictionary D. A co-occurrence vector $\mathbf{v}(x_s)$, whose each element is corresponding to a word in the source, is converted into a vector $\mathbf{v}_t(x_s)$, whose each element is corresponding to a word in the destination, referring the dictionary D:

$$\mathbf{v}_t(x_s) = (f_t(x_s, z_1), \ldots, f_t(x_s, z_m)), \qquad (2)$$

where $z_j (j = 1, 2, \ldots, m)$ is a translation word which appears in the dictionary D. The function $f_t(x_s, z_k)$, which assigns a co-occurrence degree between a word x_s and a word z_j in the destination based on a co-occurrence vector of a word x_s in the source, is defined as follows:

$$f_t(x_s, z_j) = \sum_{i=1}^{n} f(x_s, x_i) \cdot \delta(x_i, z_j). \qquad (3)$$

where $\delta(x_i, z_j)$ is equal to one when a word z_j is included in a translation word set $D(x_i)$, which consists of translation words of a word x_i, and zero otherwise.

A set of description sentences $\mathbf{Y}_s$ in the pivot are obtained referring the source-pivot dictionary for a word x_s. After that, a description sentence $\mathbf{y}_s \in \mathbf{Y}_s$ in the pivot is converted to a set of description sentences $\mathbf{Z}_s$ in the destination referring the pivot-destination dictionary. A co-occurrence vector against a candidate description sentence $\mathbf{z}_s = z_s^1 z_s^2 \cdots z_s^l$, which is an instance of $\mathbf{Z}_s$, is calculated by this equation:

$$\mathbf{u}(\mathbf{z}_s) = \left(\sum_{k=1}^{l} f(z_s^k, z_1), \ldots, \sum_{k=1}^{l} f(z_s^k, z_m) \right) \qquad (4)$$

Finally, the candidate $\mathbf{z}_s$ which meets a certain condition is selected as an output. Two conditions are examined in this paper: (1) selecting top-n candidates from sorted ones according to each similarity score, and (2) selecting candidates whose similarity scores are greater than a certain threshold. In this paper, cosine distance $s(\mathbf{v}_t(x_s), \mathbf{u}(\mathbf{z}_s))$ between a vector based on an input word x_s and a vector based on a candidate $\mathbf{z}_s$ is used as the similarity score between them.

3 Experiments

In this section, we present the experiments of the proposed method that the Indonesian language, the English language and the Japanese language are adopted as the source language, the pivot language and the destination language respectively.

3.1 Experimental Data

The proposed method depends on three translation dictionaries and two monolingual corpora as described in Section 2.

Mainichi Newspaper Corpus (1993–1995), which contains 3.5M sentences consist of 140M words, is used as the Japanese corpus. When measuring similarity between words using co-occurrence vectors, it is common that a corpus in the source language for the similar domain to one of the corpus in the source language is more suitable than one for a different domain. Unfortunately, because we could not find such corpus, the articles which were downloaded from the Indonesian Newspaper WEB sites[1] are used as the Indonesian corpus. It contains 1.3M sentences, which are tokenized into 10M words.

An online Indonesian-Japanese dictionary[2] contains 10,172 headwords, however, only 6,577 headwords of them appear in the Indonesian corpus. We divide them into two sets: the first set which consists of 6,077 entries is used as the seed dictionary, and the second set which consists of 500 entries is used to evaluate translation performance. Moreover, an online Indonesian-English dictionary[3], and an English-Japanese dictionary(Michibata, 2002) are also used as the source-pivot dictionary and the pivot-destination dictionary.

3.2 Evaluation of Translation Performance

As described in Section 2, two conditions of selecting output words among candidates are examined. Table 1 shows their performances and the baseline, that is the translation performance when all candidates are selected as output words. It is revealed that the condition of selecting top-n candidates outperforms the another condition and the baseline. The maximum $F_{\beta=1}$ value of 52.5% is achieved when selecting top-3 candidates as output words.

Table 2 shows that the lexical distribution of headwords contained in the seed dictionary are quite similar to the lexical distribution of headwords contained in the source-pivot dictionary. This observation means that it is necessary to translate verbs and adjectives as well as nouns, when expanding this seed dictionary. Table 3 shows translation performances against nouns, verbs and adjectives, when selecting top-3 candidates as output words. The proposed method can be regarded likely because it is effective to verbs and adjectives as well as to nouns, whereas the baseline precision of verbs is considerably lower than the others.

3.3 CLIR Performance Improved by Expanded Dictionary

In this section, performance impact is presented when the dictionary expanded by the proposed method is adopted to the real CLIR system proposed in (Purwarianti et al., 2007).

NTCIR3 Web Retrieval Task(Eguchi et al., 2003) provides the evaluation dataset and defines the evaluation metric. The evaluation metric consists of four MAP values: PC, PL, RC and RL. They are corresponding to assessment types respectively. The dataset consists 100GB Japanese WEB documents and 47 queries of Japanese topics. The Indonesian queries, which are manually translated from them, are used as inputs of the experiment systems. The number of unique words which occur in the queries is 301, and the number of unique words which are not contained in the Indonesian-Japanese dictionary is 106 (35%). It is reduced to 78 (26%), while the existing dictionary that contains 10,172 entries is expanded to the dictionary containing 20,457 entries with the proposed method.

Table 4 shows the MAP values achieved by both the baseline systems using the existing dictionary and ones using the expanded dictionary. The former three systems use existing dictionaries, and the latter three systems use the expanded one. The 3rd system translates keywords transitively using both

[1] http://www.kompas.com/, http://www.tempointeraktif.com/

[2] http://m1.ryu.titech.ac.jp/~indonesia/todai/dokumen/kamusjpina.pdf

[3] http://nlp.aia.bppt.go.id/kebi

Table 1: Comparison between Conditions of Selecting Output Words

	Selecting top-n candidates					Selecting plausible candidates				Baseline
	$n = 1$	$n = 2$	$n = 3$	$n = 5$	$n = 10$	$x = 0.1$	$x = 0.16$	$x = 0.2$	$x = 0.3$	
Prec.	55.4%	49.9%	46.2%	40.0%	32.2%	20.8%	23.6%	25.8%	33.0%	18.9%
Rec.	40.9%	52.6%	60.7%	67.4%	74.8%	65.3%	50.1%	40.0%	16.9%	82.5%
$F_{\beta=1}$	47.1%	51.2%	**52.5%**	50.2%	45.0%	31.6%	32.1%	31.4%	22.4%	30.8%

Table 2: Lexical Classification of Headwords

	Indonesian-Japanese	Indonesian-English
# of nouns	4085 (57.4%)	15718 (53.5%)
# of verbs	1910 (26.8%)	9600 (32.7%)
# of adjectives	795 (11.2%)	3390 (11.5%)
# of other words	330 (4.6%)	682 (2.3%)
Total	7120 (100%)	29390 (100%)

Table 3: Performance for Nouns, Verbs and Adjectives

	Noun		Verb		Adjective	
	$n = 3$	Baseline	$n = 3$	Baseline	$n = 3$	Baseline
Prec.	49.1%	21.8%	41.0%	14.7%	46.9%	26.7%
Rec.	65.6%	80.6%	52.3%	84.1%	59.4%	88.4%
$F_{\beta=1}$	56.2%	34.3%	46.0%	25.0%	52.4%	41.0%

Table 4: CLIR Performance

	PC	PL	RC	RL
(1) Existing Indonesian-Japanese dictionary	0.044	0.044	0.037	0.037
(2) Existing Indonesian-Japanese dictionary and Japanese proper name dictionary	0.054	0.052	0.047	0.045
(3) Indonesian-English-Japanese transitive translation with statistic filtering	0.078	0.072	0.055	0.053
(4) Expanded Indonesian-Japanese dictionary	0.061	0.059	0.046	0.046
(5) Expanded Indonesian-Japanese dictionary with Japanese proper name dictionary	0.066	0.063	0.049	0.049
(6) Expanded Indonesian-Japanese dictionary with Japanese proper name dictionary and statistic filtering	0.074	0.072	0.059	0.058

the source-pivot dictionary and the pivot-destination dictionary, and the others translate keywords using either the existing source-destination dictionary or the expanded one. The 3rd system and the 6th system try to eliminate unnecessary translations based statistic measures calculated from retrieved documents. These measures are effective as shown in (Purwarianti et al., 2007), but, consume a high run-time computational cost to reduce enormous translation candidates statistically. It is revealed that CLIR systems using the expanded dictionary outperform ones using the existing dictionary without statistic filtering. And more, it shows that ones using the expanded dictionary without statistic filtering achieve near performance to the 3rd system without paying a high run-time computational cost. Once it is paid, the 6th system achieves almost same score of the 3rd system. These observation leads that we can conclude that our proposed method to expand dictionary is valuable to a real CLIR system.

4 Concluding Remarks

In this paper, a novel method of expanding a small existing translation dictionary to a large translation dictionary using a pivot language is proposed. Our method uses information obtained from a small existing translation dictionary from the source language to the destination language effectively. Experiments that expands the Indonesian-Japanese dictionary using the English language as a pivot language shows that the proposed method can improve performance of a real CLIR system.

References

Koji Eguchi, Keizo Oyama, Emi Ishida, Noriko Kando, , and Kazuko Kuriyama. 2003. Overview of the web retrieval task at the third NTCIR workshop. In *Proceedings of the Third NTCIR Workshop on research in Information Retrieval, Automatic Text Summarization and Question Answering*.

Hideki Michibata, editor. 2002. *Eijiro*. ALC, 3. (in Japanese).

Ayu Purwarianti, Masatoshi Tsuchiya, and Seiichi Nakagawa. 2007. Indonesian-Japanese transitive translation using English for CLIR. *Journal of Natural Language Processing*, 14(2), Apr.

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th International Conference on Computational Linguistics*.

Masatugu Tonoike, Mitsuhiro Kida, Toshihiro Takagi, Yasuhiro Sasaki, Takehito Utsuro, and Satoshi Sato. 2005. Translation estimation for technical terms using corpus collected from the web. In *Proceedings of the Pacific Association for Computational Linguistics*, pages 325–331, August.