Computational Linguistics and Chinese Language Processing
Vol. 10, No. 3, September 2005, pp. 329-346

Collocational Translation Memory Extraction Based on Statistical and Linguistic Information

Thomas C. Chuang[*], Jia-Yan Jian[+], Yu-Chia Chang[+] and Jason S. Chang[+]

Abstract

In this paper, we propose a new method for extracting bilingual collocations from a parallel corpus to provide phrasal translation memories. The method integrates statistical and linguistic information to achieve effective extraction of bilingual collocations. The linguistic information includes parts of speech, chunks, and clauses. The method involves first obtaining an extended list of English collocations from a very large monolingual corpus, then identifying the collocations in a parallel corpus, and finally extracting translation equivalents of the collocations based on word alignment information. Experimental results indicate that phrasal translation memories can be effectively used for computer assisted language learning (CALL) and computer assisted translation (CAT).

Keywords: Bilingual Collocation Extraction, Collocational Translation Memory, Collocational Concordancer

1. Introduction

Example-based machine translation (EBMT) has been proposed as an alternative approach to automatic translation. Translations of examples range from two-word to multi-word, translations, with or without syntactic or semantic structures [Nagao 1984; Kitano 1993; Smadja 1993; Lin 1998; Andrimanankasian *et al*. 1999; Carl 1999; Brown 2000; Pearce 2001; Seretan *et al.* 2003]. In the approach, text and translations are preprocessed and stored in a translation memory, which serves as an archive of existing translations that the MT system can reuse. A number of proposed applications for machine translation and computer assisted translation systems use translation examples found in bilingual corpora; these methods include

[*] Department of Computer Science and Information Engineering, Vanung University, No. 1, Vanung Road, Jhongli, Taoyuan, Taiwan
E-mail: tomchuang@msa.vnu.edu.tw

[+] Department of Computer Science, National Tsing Hua University, 101, Kuangfu Road, Hsinchu, Taiwan

[Received January 31, 2005; Revised August 24, 2005; Accepted August 26, 2005]

[Transit 2005], [Deja–Vu 2005], [TransSearch 2005], and [TOTALrecall 2005].

Statistical methods have been proposed for automatic acquisition of bilingual collocations [Smadja *et al.* 1996; Gao *et al*. 2002; Wu and Zhou 2003] from parallel bilingual corpora [Kupiec 1993; Smadja *et al.* 1996; Echizen-ya *et al.* 2003] or from two comparable monolingual corpora [Lu and Zhou 2004]. These bilingual collocations, if acquired in quantity, can enable a machine translation system to produce more native-speaker like translations. However, parallel corpora of substantial size are harder than monolingual corpora to come by. Therefore, in small- to mid-size parallel texts, collocations may not have high enough counts for a statistical method to reliably extract them.

Consider the example of extracting verb-noun collocations for the noun “influence” from 50,000 bilingual sentences (SMEC-50000) in the Sinorama Mandarin-English Corpus (SMEC)[1]. Some useful bilingual collocations in SMEC have very low occurrence counts. For instance, the bilingual collocation “use influence; 發揮 影響力” appears only once in SM-50000 (see Example 1). Such collocations may not be extracted by the methods proposed in the literature.

(1) These circumstances make it unlikely that APEC will be able to avoid reform. In Lai's analysis, the implosion of the Asian economies last year demonstrates their interconnectedness. Therefore, in order to place its own existence on a more secure foundation, Taiwan should carefully observe changes in APEC and **use** its **influence** to make the organization into a vehicle driving regional consolidation.
他分析，亞洲國家近年來經濟危機持續不去，證明瞭彼此的連動性，因此台灣應該注意觀察APEC的轉變，發揮意見的影響力，以使APEC能夠成爲區域整合的火車頭，爲我國創造更大的生存利基。

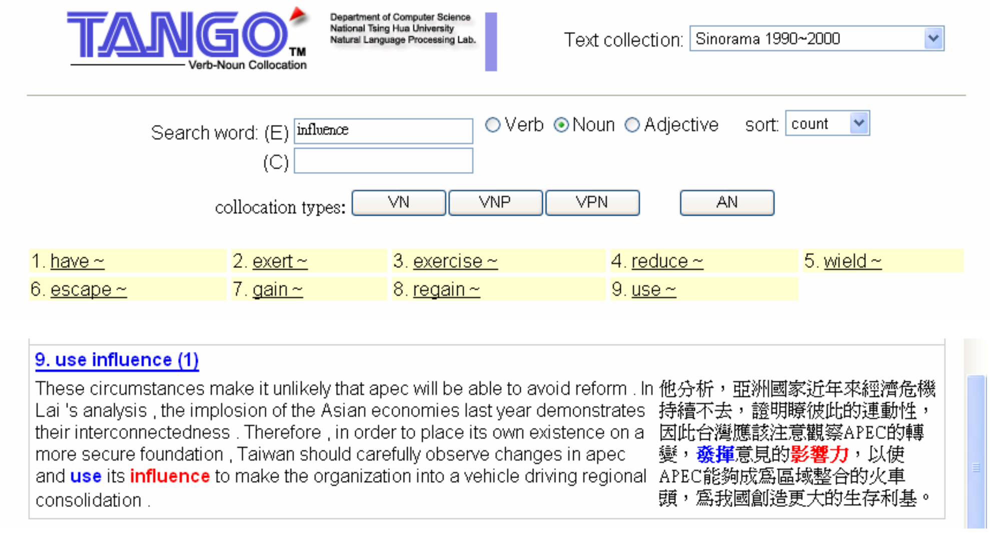
A good way of extracting such bilingual collocation might be to first extract “use influence” as a collocation in a large, separate, monolingual corpus, and then identify its instances and translations in the given parallel corpus (e.g., the Sinorama Mandarin-English Corpus). At present, it is not difficult to obtain a much larger monolingual corpus (e.g., the British National Corpus) that contains enough instances of “use influence” such that extraction of such a collocation type is mostly effective. Example (2) shows one of the 60 instances of

[1] The Sinorama Mandarin-English Corpus was originally a database of some 6,000 bilingual articles appearing in Sinorama Magazine dated 1976 to 2001 (Copyright ©2001 Sinorama Magazine & Wordpedia.com Co.) The database is a parallel text corpus, now available from The Association for Computational Linguistics and Chinese Language Processing.

“use influence” in BNC. Such a relatively high count makes it very easy to identify “use influence” as a collocation by using any the method proposed in the literature.

(2) I don't know who it and apparently he asked him that, are, are any of your men gonna be there and if there are he said, I'm, I'm gonna pull out and **use** all my **influence** to stop the march and the IRA police said no there would not be any gunmen there so I thought yeah, fucking right, oh yeah that's easy to say, and then if like the reporter said and, and you believe him and you have the feeble excuse towards a small community he said, you know what 's going on.

We will present a new method that automatically performs shallow parsing on an English corpus to identify all the statistically significant collocation types and their instances in the monolingual corpus and the English part of the given bilingual corpus. After that, the translation of each collocate of each collocation is identified based using primarily the word alignment technique. We will also present a computer assisted translation system, *Tango*, which accepts user queries of words, parts of speech, and types of collocation, and displays citations with bilingual collocations highlighted. An example of a Tango search for bilingual collocations for the noun “influence” is shown in Figure 1.


TANGO™
Verb-Noun Collocation
Department of Computer Science
National Tsing Hua University
Natural Language Processing Lab.

Text collection: Sinorama 1990~2000

Search word: (E) influence

(C)

○ Verb ● Noun ○ Adjective

sort: count

collocation types:
VN
VNP
VPN
AN

1. have ~ 2. exert ~ 3. exercise ~ 4. reduce ~ 5. wield ~
6. escape ~ 7. gain ~ 8. regain ~ 9. use ~

9. use influence (1)

These circumstances make it unlikely that apec will be able to avoid reform . In Lai 's analysis , the implosion of the Asian economies last year demonstrates their interconnectedness . Therefore , in order to place its own existence on a more secure foundation , Taiwan should carefully observe changes in apec and **use** its **influence** to make the organization into a vehicle driving regional consolidation .

他分析，亞洲國家近年來經濟危機持續不去，證明瞭彼此的連動性，因此台灣應該注意觀察APEC的轉變，**發揮**意見的**影響力**，以使APEC能夠成爲區域整合的火車頭，爲我國創造更大的生存利基。

Figure 1. An example of a Tango search for bilingual collocations for the noun “influence”.

The rest of the article is organized as follows. We will review related works in the next section. Then we will present our method for automatically processing sentences in monolingual and parallel corpora, and extracting bilingual collocations (Section 3). As part of our evaluation, we will describe an implementation of the proposed approach using SMEC and BNC (Section 4) and discuss the results of our evaluation carried out to assess the performance of bilingual collocation extraction (Section 5).

2. Extraction of Collocational Translation Memory

It is difficult to extract bilingual collocations from parallel corpora due to their limited availability and small sizes. Methods proposed in previous works typically extract collocations based solely on co-occurrence counts and statistical association measures in bilingual corpora. Unfortunately, a substantial part of the collocations in a modest-sized parallel corpus might not have high enough frequency counts for statistical extraction methods to be effective. Many bilingual collocations useful for machine translation and computer assisted language learning may appear only once or twice in a small to medium size corpus. To extract bilingual collocations, a promising approach is to acquire collocations in from a very large monolingual corpus and obtain translations from a parallel corpus.

2.1 Problem Statement

We will focus here on the first step of building a translation memory for a bilingual collocation tool this involves; extracting a set of bilingual collocations instances from a sentence-aligned parallel corpus. These collocation instances can be used for the purpose of computer assisted translation and language learning. Thus, the collocations, including those that appear only once or twice, should be identified in the part of the parallel corpus that is written in one of the two languages. At the same time, it is crucial that the translations also be identified. Therefore, our goal is to return a reasonable-size set of documents that, at the same time, must contain an answer to the question. For simplicity, we will focus on verb-noun collocations in this paper. We formally state the problem that we are addressing below.

Problem Statement: Given a parallel corpus *PC* of *n* pairs of sentences (SE_i, SF_i) written in the first language *E* and the second language *F*. Our goal is to identify a set of *k collocations* and *translations* (CE_{ij}, CF_{ij}) in (SE_i, SF_i). To accomplish this task, we make use of a large corpus *M* with *m* sentences ME_i of texts written in *E* to help identify CE_{ij} in SE_i.

We attempt to identity collocations in a parallel corpus by leveraging another larger monolingual corpus in which collocations appears with higher occurrence counts. Our method is shown in Figure 2.

(1)	Annotate the English Sentences with parts of speech, chunk, and clause information (Section 2.2)
(2)	Extract English collocation types in ME_{i} (Section 2.3)
(3)	Extract English collocation instances in CE_{ij} in SE_{i} (Section 2.3)
(4)	Identify translation equivalents CF_{ij} to collocation CE_{ij} in SF_{i} (Section 2.4)

Figure 2. Outline of the process used to extract bilingual collocations

2.2 Taggers for parts of speech, chunks, and clauses

Using an annotated corpus with texts written in *E*, we can develop a tagger based on three Hidden Markov Models: one each for parts of speech, chunks (i.e. basis phrases), and clauses. The training corpus consists of sentences with three levels of annotation for each word: parts of speech, chunk, and clause. Figure 3 shows three levels of annotation for Example (3):

w_{i}	t_{i}	u_{i}	v_{i}
This	DT	B–NP	S
is	VBZ	B–VP	X
the	DT	B–NP	X
11th	JJ	I–NP	X
consecutive	JJ	I–NP	X
quarter	NN	I–NP	X
in	IN	B–PP	S
which	WDT	B–NP	X
the	DT	B–NP	S
company	NN	I–NP	X
has	VBZ	B–VP	X
paid	VBN	I–VP	X
shareholders	NNS	B–NP	X
an	DT	B–NP	X
extra	JJ	I–NP	X
dividend	NN	I–NP	X
of	IN	B–PP	X
five	CD	B–NP	X
cents	NNS	I–NP	X
.	.	O	X

Figure 3. Examples of three levels of tagging performed on the sentence "This is the 11th consecutive quarter in which the company has paid shareholders an extra dividend of five cents."

(3) This is the 11th consecutive quarter in which the company has paid shareholders an extra dividend of five cents.

As shown in Figure 3, each word is tagged with a parts of speech tag (e.g., DT for determiner and VBZ for third person singular verb), a chunk tag indicating the basis phrase type (e.g., noun phrase, NP, verb phrase VP, etc.) Plus the position of the word (e.g., "B" for words beginning a chunk and "I" for all other words in the chunk), and a clause tag (e.g., "S" for a clause-beginning word and "X" otherwise). The chunk annotation shown in Figure 3 indicates that "This," "the 11th consecutive quarter," "the company," "shareholder," "an extra dividend," and "five cents" are nouns on noun phrases, while "is" and "has paid" are verb phrases. Similarly, the clause annotation results indicate that "This is the 11th consecutive quarter" and "in which the company has paid shareholders an extra dividend of five cents" are the only two clauses in the sentence. With a substantial amount of annotated sentences like the above we can develop three taggers for each level of analysis.

For the parts of speech tagger, the HMM operates on a set of states represented by all possible POS tage, goes through a sequence of state $t_{i,}$ and produces words w_i, i = 1 to n. An example of transition probability function $\mathrm{P}(u_i \mid u_{i-1})$ for the chunk tagger is shown in Figure 4. A first order HMM is characterized by the emission probability, $\mathrm{P}(w_i \mid t_i)$, and state transition probability, $\mathrm{P}(t_i \mid t_{i-1})$. An example of the emission probability function $\mathrm{P}(t_i \mid u_i)$ for the chunk tagger is shown in Figure 5.

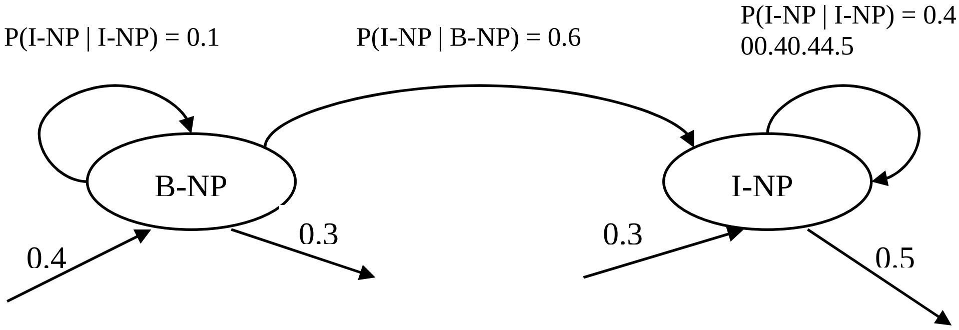


Figure 4. Example of transition probability function $\mathrm{P}(u_i \mid u_{i-1})$ ***for the chunk tagger***

POS / state	DT	NN	Others
B-NP	0.6	0.2	0.2
I-NP	0.1	0.7	0.2
others	…	…	…

Figure 5. Example of the emission probability function $\mathrm{P}(t_i \mid u_i)$ ***for the chunk tagger***

Specifically, we have

$$P(w) = P(w_1 w_{2 \dots} w_n) = P(t_1)\, P(w_1 \mid t_1) \prod_{i=2,n} P(t_i \mid t_{i-1})\, P(w_i \mid t_i)\ . \qquad (1)$$

Therefore, we can to derive the parts of speech $t = (t_1\, t_{2 \dots}\, t_n)$ for the sentence $w = (w_1\, w_{2 \dots}\, w_n)$ by calculating $(t_1\, t_{2 \dots}\, t_n)$ that maximizes $P(t \mid w)$. Thus, we have

$$(t_1\, t_{2 \dots}\, t_n) = \operatorname{argmax}_t P(t \mid w) = \operatorname{argmax}_t P(w \mid t)\, P(t)/P(w) = \operatorname{argmax}_t P(w \mid t)\, P(t)$$
$$= \operatorname{argmax}_t P(t_1)\, P(w_1 \mid t_1) \prod_{i=2,n} P(t_i \mid t_{i-1})\, P(w_i \mid t_i)\ . \qquad (2)$$

Similarly, we can derive the chunk tags $(u_1\, u_{2 \dots}\, u_n)$ and clause tags $(v_1\, v_{2 \dots}\, v_n)$ for the given sentence $(w_1\, w_{2 \dots}\, w_n)$ using Equations (3) and (4):

$$(u_1\, u_{2 \dots}\, u_n) = \operatorname{argmax}_u P(u_1)\, P(t_1 \mid u_1) \prod_{i=2,n} P(u_i \mid u_{i-1})\, P(t_i \mid u_i)\ , \qquad (3)$$

$$(v_1\, v_{2 \dots}\, v_n) = \operatorname{argmax}_v P(v_1)\, P(u_1 \mid v_1) \prod_{i=2,n} P(v_i \mid v_{i-1})\, P(u_i \mid v_i)\ . \qquad (4)$$

The optimal values of parts of speech tags $(t_1\, t_{2 \dots}\, t_n)$, chunk tags $(u_1\, u_{2 \dots}\, u_n)$, and clause tags $(v_1\, v_{2 \dots}\, v_n)$ can be derived by using a dynamic procedure [Manning and Shutze 1999]. The tagging process is carried out for the *n* source sentences SE_i in a given parallel corpus *PC* and in the *m* sentences ME_i in a large corpus *M*. We will describe in Section 3 how the training data provided the results from the common task CoNLL-2000 and CoNLL-2001 shared tasks (CoNLL, 2000) can be used to estimate the probabilistic functions involved, including $P(t_1)$, $P(t_i \mid t_{i-1})$, $P(w_i \mid t_i)$, $P(u_1)$, $P(u_i \mid u_{i-1})$, $P(t_i \mid u_i)$, $P(v_i)$, $P(v_i \mid v_{i-1})$ and $P(u_i \mid v_i)$.

2.3 Extraction of Collocation Types in *M*

With the chunk and clause tags for the sentences in the monolingual corpus *M*, we can proceed to extract a set of verb-noun collocation types from *M*. To that end, we can consider the heads of phrases in three prevalent verb-noun collocation structures in the corpus: VP+NP, VP+PP+NP, and VP+NP+PP. To extract verbs and nouns that appear in a predicate-object relation, we need to have a full parse of the sentences. However, a state-of-the-art parser can not produce a full parse of unrestricted texts with a very high precision rate. Therefore, we simply assume that a noun phrase following a verb phrase is in a predicate-object relationship unless they belong to two different clauses.

For instance, consider the sentence shown in Example (4):

(4) Confidence in the pound is widely expected to take another sharp dive if trade figures for September due for release tomorrow fail to show a substantial improvement from July and August’s near-record deficits.

The taggers described in Section 2.2 will produce the parts of speech tags, chunk tags, and clause tags shown in Examples (5)-(7):

(5) Confidence/NN in/IN the/DT pound/NN is/VBZ widely/RB expected/VBN
to/TO take/VB another/DT sharp/JJ dive/NN if/IN trade/NN figures/NNS
for/IN September/NNP ,/, due/JJ for/IN release/NN tomorrow/NN ,/,
fail/VB to/TO show/VB a/DT substantial/JJ improvement/NN from/IN
July/NNP and/CC August/NNP 's/POS near-record/JJ deficits/NNS ./.

(6) Confidence /B-NP in/B-PP the/B-NP pound/I-NP is/B-VP widely/I-VP expected/I-VP
to/I-VP take/I-VP another/B-NP sharp/I-NP dive/I-NP if/B-SBAR trade/B-NP figures/I-NP
for/B-PP September/B-NP due/ADJP for/B-PP release/B-NP tomorrow/I-NP ,/O
fail/O to/O show/O a/B-NP substantial substantial/I-NP improvement/I-NP from/B-PP
July/B-NP and/O August/B-NP 's/B-NP near-record/I-NP deficits/I-NP ./O

(7) Confidence /S in/X the/X pound/X is/X widely/X expected/X
to/S take/X another/X sharp/X dive/X if/S trade/X figures/X
for/X September/X due/X for/X release/X tomorrow/X ,/ *
fail/S to/X show/X a/X substantial/X improvement/X from/X
July/X and/X August/X 's/X near-record/X deficits/X ./X

The words in the same chunk can be further grouped together (as shown in Figure 6) to make it easy to examine the phrase types of two adjacent chunks and extract the head word of each phrase. For instance, we can extract a VN pair (e.g., “*take*” and “*dive*”) from an annotated sentence by taking the last words of two adjacent VP and NP chunks.

Phrase	Type
Confidence	NP
In	PP
the pound	NP
is widely expected to ***take***	VP
another sharp ***dive***	NP
If	SBAR
trade figures	NP
For	PP
September	NP

Figure 6. Example of grouping words in a chunk together record by record

Care should be taken to avoid extracting verbs and nouns from two clauses in a sentence. For instance, in Example (8), in some cases, considering only chunk information is not sufficient. For example, the VN pair ("think," "people") taken from the two separate clauses "why do you think" and "people cannot see the top of the building on some days" should be excluded from consideration.

(8)Why do you think people cannot see the top of the building on some days?

Some VN pairs extracted at this stage are free combinations, while some recur more frequently than is likely according to chance and should be considered collocations. After obtaining a list of instances of candidate collocations, we proceed to find distinct collocation types and calculate their word counts and phrase counts in order to verify whether each of them is a valid collocation. After that, we calculate the strength of association between each verb-noun pair in the collocations by using Logarithmic Likelihood Ratio (LLR) statistics. Equation (5) is the formula that computes the LLR.

$$LLR(x;y) = -2\log_2 \frac{p_1^{k_1}(1-p_1)^{n_1-k_1}\,(1-p_2)^{n_2-k_2}}{p^{k_1}(1-p)^{n_1-k_1}\;p^{k_2}(1-p)^{n_2-k_2}}; \tag{5}$$

k_1 : count of sentences that contain x and y simultaneously;

k_2 : count of sentences that contain x but do not contain y;

n_1 : count of sentences that contain y;

n_2 : count of sentences that do not contain y;

$p_1 = k_1 / n_{1;}$

$p_2 = k_2 / n_{2;}$

$p = (k_1+k_2) / (n_1+n_2)$.

We have described a method for handling VN collocations. This method can be easily extended to handle VPN and VNP collocations as well. The idea is quite simple. After identifying a VN collocation type where the verb and noun are separated by a preposition, we go on to consider the preposition that comes between the verb and the noun or that follows the verb and noun. The VPN and VNP collocations are validated again by calculating the LLR between each VN pair and the preposition.

2.4 Extraction of Collocation Instances in *PC*

We subsequently identify collocation instances in the *n* sentence SE_i of the give parallel corpus *PC*. First, each sentence SE_i is subjected to the same POS, chunk, and clause analyses as is applied to the corpus *M*. The collocation instances of the forms VN, VPN, and VNP are extracted in a similar way to that described in Section 2.2. There are two cases in which a collocation instance will be considered as a valid collocation:

1. if it passes the LLR threshold calculated based on the counts of words and co-occurrences in *PC*;
2. if it is in the list of valid collocations found in *M*.

The quantity and quality of collocations in a very large monolingual corpus surely will facilitate collocation identification in a smaller bilingual corpus with better statistical measures.

2.5 Extracting Collocation Translation Equivalents in a Bilingual Corpus

After instances that are most likely valid collocations are obtained from a bilingual corpus, we go on to work on the second part of the parallel corpus *PC*. We exploit statistical word-alignment techniques [Melamed 1997] and dictionaries to find translation candidates for each of the words in a given collocation. Using Melamed’s approach, we can establish a word translation based on corpora to supplement English-Chinese dictionaries, which generally suffer due to insufficient information. We first locate the translation of the noun. Subsequently, we locate the verb nearest to the noun translation to find the translation of the verb. Figure 7 shows some examples.

English sentence	Chinese sentence
If in this time no one shows concern for them, and directs them to correct thinking, and teaches them how to express and ***release emotions***, this could very easily leave them with a terrible personality complex	如果這時沒有人關心他們,引導他們正確思考，教他們表達、渲洩**情緒**，極易在人格成長上留下一個打不開的死結。

they can never resolve.	
Occasionally some kungfu movies may appeal to foreign **audiences**, but these too are exceptions to the rule.	偶爾有一些武打片對某些外國**觀眾**有吸引力，但也是個案。

Figure 7. Examples of identifying translations of nouns (in bold) and verbs (shaded) of VN collocation instances in bilingual sentence pairs

3. Implementation and evaluation

We have implemented a program for extracting bilingual collocations based on the proposed method and experimented with 50,000 bilingual sentences (SMEC-50000) from the Sinorama Mandarin-English Corpus (SMEC). We wanted to assess the performance of the program and verify whether useful bilingual collocations in SMEC with very low occurrence counts (e.g., “use influence; 發揮 影響力”) could be extracted. Such collocations are beyond the reach of methods previously proposed in the literature.

We used the Brown corpus to develop a parts-of-speech tagger and the CoNLL-2000 benchmark database to build a chunk tagger and clause tagger. The chunk tagger relied on the transition and output probabilities of chunks. Figures 8 and 9 show examples of these two processes. The average precision rate of the chunk tagger was about 93.7%, based on CoNLL testing data.

Chunk tag u_i	Chunk tag u_{i+1}	adj. count(u_i u_{i+1})	count(u_i)	P(u_i | u_{i+1})
B-NP	I-NP	46327.3	67503	0.686300
B-NP	B-VP	8762.3	67503	0.129806
B-NP	O	5418.3	67503	0.080268
B-NP	B-PP	3878.3	67503	0.057454
B-NP	B-NP	1974.3	67503	0.029248
B-NP	B-ADVP	645.3	67503	0.009560
B-VP	I-VP	9830.3	26125	0.345313
B-VP	B-NP	9021.3	26125	0.097619
B-VP	B-PP	2550.3	26125	0.065811
B-VP	O	1719.3	26125	0.039782
B-VP	B-ADJP	1039.3	26125	0.031169
B-VP	B-ADVP	814.3	26125	0.025428
B-VP	B-SBAR	664.3	26125	0.011265

Figure 8. Example data of transition probabilities of chunks

Chunk tag u_i	POS tag t_i	adj. count($u_i\, u_{i+1}$)	count(u_i)	P($t_i \mid u_i$)
B-NP	at	19307.3	67503	0.286021
B-NP	nn	7555.3	67503	0.111925
B-NP	np	7541.3	67503	0.111718
B-NP	jj	5587.3	67503	0.082771
B-NP	nns	4473.3	67503	0.066268
B-NP	cd	2836.3	67503	0.042017
B-NP	pp$	2261.3	67503	0.033499
B-NP	pps	2080.3	67503	0.030818
B-NP	ppss	1620.3	67503	0.024003
I-NP	nn	34671.3	77683	0.446318
I-NP	nns	13143.3	77683	0.169191
I-NP	jj	8247.3	77683	0.106166
I-NP	np	7250.3	77683	0.093332
I-NP	cd	4727.3	77683	0.060854
I-NP	cc	1727.3	77683	0.022235
I-NP	vbg	1147.3	77683	0.014769
I-NP	vbn	940.3	77683	0.012104
I-NP	ap	862.3	77683	0.011100

Figure 9. Example data of emission probabilities of chunks

Using the chunk and clause information, we proceeded to extract a list of collocation types from the monolingual British National Corpus. We mainly used this list to identify collocation instances in SMEC. Finally, we applied the Competitive Linking Algorithm to SMEC to obtain word alignment results. We then applied the results of word alignment to extract the matching translations of the noun and verb collocates. The collocation extraction program produced a much larger set of collocation candidates than could be obtained from BNC. The corpus consists of over 100 million words in about 5 million sentences. After filtering out incomplete sentences, we obtained around 4 million sentences for use in extracting valid English collocations. After implementing our proposed method as described in Sections 2.2 and 2.3, we obtained over half a million collocation types of the forms VN, VPN, and VNP. We were able to identify over 30,000 collocation instances in SMEC. Figures 10 and 11 show some examples in BNC.

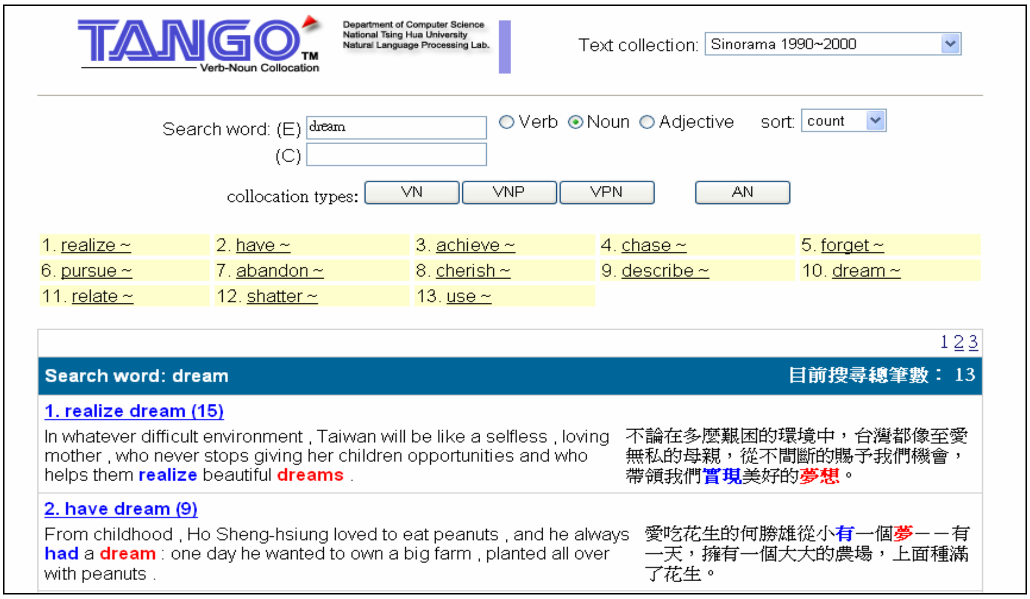
Type	Collocation types in the British Nation Corpus (BNC)	Collocation instances in the Sinorama Parallel Corpus (SPC)
VN	631,638	26,315
VPN	15,394	3,457
VNP	14,008	4,406

Figure 10. The results for collocation types extracted from the BNC and SMEC

VN type	Example
Exert influence	That means they would already be exerting their influence by the time the microwave background was born.
Exercise influence	The Davies brothers, Adrian (who scored 14 points) and Graham (four), exercised an important creative influence on Cambridge fortunes while their flankers Holmes and Pool-Jones were full of fire and tenacity in the loose.
Wield influence	Fortunately, George V had worked well with his father and knew the nature of the current political trends, but he did not wield the same influence internationally as his esteemed father.
Extend influence	The CAB extended its influence into the non-government sector, funding research by the Cathedral Advisory Commission and the Royal Society for the Protection of Birds.
Diminish influence	To break up the Union now would diminish our influence for good in the world, just at the time when it is most needed.
Gain influence	In general, women have not benefited much in the job market from capitalist industrialization nor have they gained much influence in society outside the family through political channels.
Counteract influence	To try and counteract the influence of the extremists, the moderate wing of the party launched a Labour Solidarity Campaign in 1981.
Reduce influence	Whether the curbs on police investigation will reduce police influence on the outcome of the criminal process is not easy to determine.

Figure 11. Examples of collocation instances extracted from SMEC

With the collocation types and instances extracted from the corpus, we built an on-line collocation reference tool called TANGO to support searching for collocations and translations of a given word.


TANGO™
Verb-Noun Collocation
Department of Computer Science
National Tsing Hua University
Natural Language Processing Lab.
collocation types:
VN
VNP
VPN
AN

Text collection: Sinorama 1990~2000

Search word: (E) dream
(C)

○ Verb ● Noun ○ Adjective sort: count

1. realize ~	2. have ~	3. achieve ~	4. chase ~	5. forget ~
6. pursue ~	7. abandon ~	8. cherish ~	9. describe ~	10. dream ~
11. relate ~	12. shatter ~	13. use ~		

1 2 3

Search word: dream 目前搜尋總筆數： 13

1. realize dream (15)

In whatever difficult environment , Taiwan will be like a selfless , loving mother , who never stops giving her children opportunities and who helps them **realize** beautiful **dreams** .

不論在多麼艱困的環境中，台灣都像至愛無私的母親，從不間斷的賜予我們機會，帶領我們**實現**美好的**夢想**。

2. have dream (9)

From childhood , Ho Sheng-hsiung loved to eat peanuts , and he always **had** a **dream** : one day he wanted to own a big farm , planted all over with peanuts .

愛吃花生的何勝雄從小**有**一個**夢**——有一天，擁有一個大大的農場，上面種滿了花生。

Figure 12. TANGO, a web-based bilingual collocation tool

TANGO accepts a query word in English and a collocation type, and returns a list of collocation types and examples. Figure 12 shows a screen returned for a query for VN collocations of “influence.” One instance for each collocation type is shown first. All instances can be shown on demand. Besides showing bilingual collocation extractions, TANGO also color codes the translation counterparts of the collocation instances. This informative, bilingual reference tool has been used in language learning classes and by professional translators. Initial responses have been quite positive, indicating that this new tool is very useful for EFL learners and translators.

To assess the quality of the extracted bilingual collocations, we randomly selected 100 sentences with extracted bilingual collocations from SMEC for manual evaluation. Many of these sentence had more than one collocation, 50 we evaluate each collocation individually. Students majoring in English assessed each bilingual collocation in the context of the corresponding pair of sentences. The evaluation process involved judging the validity of translation of the collocation. There were three levels of validity: satisfactory translation, approximate translation (partial matching), and unacceptable translation. Figure 13 shows examples for each level of validation. For the purpose of this research, satisfactory translations and approximate translations were considered useful. Therefore, we determined the percentages of bilingual collocations that fell into these two categories. As indicated in Figure 14, the average precision rate for the extracted bilingual collocations was about 90% for satisfactory translations and approximate translations.

Level of quality	English sentences	Chinese sentences
satisfactory translation	Thus when Chinpao Shan put out its advertisement last year, looking for new people to develop its related **enterprises**, the notice frankly stated "Southern Taiwanese preferred.	去年，金寶山在發展**關係企業**徵招新人的廣告上，就坦白指明「本省籍南部人優先」。
approximant translation	Ah-ying relates that "Teacher Chang" friendly and easy-going, is always there to answer her **questions**. She even goes to him for answers when her friends have legal questions.	阿英表示，「張老師」親切隨和，只要有不懂的事，都去問老師，就連朋友有法律上的**問題**，也去請教他。
unacceptable translation	Said one observer, "If I can speak bluntly, the mainlanders are robbing graves of their treasures and smuggling them away, and the situation is bad. In reality, though, it is Taiwan that is behind it all committing the **crime**.	「說得不好聽，大陸近年來盜墓、文物走私情形嚴重，台灣其實是背後的劊子手！」有人這樣**認爲**。

Figure 13. Three levels of quality of the extracted translation memory

Type	% of satisfactory translations	% of satisfactory and approximate translations*
VN	73	90
VPN	66	89
VNP	78	89

Figure 14. Evaluation of bilingual collocations extracted from SMEC

4. Discussion and limitation

Collocation is an important part of translation task yet it has long been neglected. Traditional machine translation tends to translate input texts word by word, which easily leads to literal translations. Therefore, even with abundant vocabulary, dictionary and grammar rule-based model systems fail to generate fluent translations in a target language. For example, due to its lack of collocational knowledge, a machine translation system may recognize “take” as “na” (i.e., take away) and “medicine” as “yao” (i.e., medicine) in Chinese, respectively. Thus, systems are inclined to literally translate “take medicine” as “na yao” (i.e., "take away the medicine" in Chinese), resulting in an odd translation or mistranslation. We suggest that machine translation systems should take collocational translation memory into consideration to improve the translation quality.

Due to the limitations of the word-alignment technique, our method may incorrectly recognize some matching translations. We need better word-alignment to align translations more correctly. Moreover, expansion of the bilingual corpora would also increase the precision achieved in retrieving collocational translation memory. This would enable us to obtain high enough counts for each collocate (i.e., verbs and nouns in VN collocations) in the target language so as to increase the confidence level of the LLR statistics, which in turn would eliminate the anomalous collocational translation memory.

5. Conclusion

In the field of machine translation, Example-Based Machine Translation (EBMT) exploits existing translations in the hope of producing better quality translations. However, collocational translation has always been neglected and is hard to deal with. We have proposed the use of collocational translation memory to develop a better translation method that can solve some problems resulting from literal translation. Encouraged by the satisfactory precision rates in collocation and translation extraction obtained in this study, we hope that collocational translation memory can be further applied in machine translation, cross language information retrieval, computer assisted language learning, and other NLP applications.

Acknowledgements

This work was part of the "CANDLE" project funded by the National Science Council of Taiwan (NSC-2213-E-238-015 and NSC92-2524-S007-002). Further information about CANDLE is available at http://candle.cs.nthu.edu.tw/.

References

Andriamanankasina, T., K. Araki, and T. Tochinai, "Example-Based Machine Translation of Parts-Of-Speech Tagged Sentences by Recursive Division," In *Proceedings of MT SUMMIT VII,* Singapore, 1999.

Brown, R. D., "Automated Generalization of Translation Examples," In *Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000),* Saarbrücken, Germany, August 2000, pp. 125-131.

Carl, M., "Inducing Translation Templates for Example-Based Machine Translation," In *Proceedings of MT Summit VII*, 1999.

CoNLL yearly meeting of the SIGNLL, the Special Interest Group on Natural Language Learning of the Association for Computational Linguistics. The shared task of text chunking in CoNLL-2000 is available at http://cnts.uia.ac.be/conll2000/.

Echizen-ya, H., K. Araki, Y. Momouchi, and K. Tochinai, "Effectiveness of Automatic Extraction of Bilingual Collocations Using Recursive Chain-Link-Type Learning," *The 9th Machine Translation Summit,* 2003, pp.1 02-109.

Gao, J., J. Nie, H. He, W. Chen, and M. Zhou,"Resolving Query Translation Ambiguity Using a Decaying Co-occurrence Model and Syntactic Dependence Relations," *The 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 2000, pp.183 -190.

Kitano, H., "A Comprehensive and PracticalModel of Memory-Based Machine Translation," In *Proceedings of IJCAI-93,* 1993, pp. 1276-1282.

Kupiec, J., "An algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora," In *Proceedings of the 31th Annual Meeting of Association for Computational Linguistics*, 1993, pp. 17-22.

Lin, D., "Extracting Collocation from Text Corpora," *First Workshop on Computational Terminology*, 1998, pp. 57-63.

Lü, Y., and M. Zhou, "Collocation Translation Acquisition Using Monolingual Corpora," *Association for Computational Linguistics* 2004, pp. 167-174.

Melamed, I. D., "A Word-to-Word Model of Translational Equivalence," In *Proceedings of the Association for Computational Linguistics* 1997, *Madrid Spain,* 1997, pp. 490-497.

Nagao, M., "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle," *Artificial and Human Intelligence,* A. Elithorn and R. Banerji (eds.) North-Holland, 1984, pp. 173-180.

Pearce, D., "Synonymy in Collocation Extraction," In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, CMU*, 2001.

Seretan, V., L. Nerima, and E. Wehrli, "Extraction of Multi-Word Collocations Using Syntactic Bigram Composition," *International Conference on Recent Advances in NLP*, 2003, pp. 424-431.

Manning, C.D., and H. Schutze, "Foundations of Statistical Natural Language Processing" *MIT Press, Cambridge, Mass*, 1999.

Smadja, F., "Retrieving Collocations from Text: Xtract," *Computational Linguistics*, 1993, 19(1), pp.143-177.

Smadja, F., K. R. Mckeown, and V. Hatzivassiloglou, "Translation Collocations for Bilingual Lexicons: a Statistical Approach," *Computational Linguistics*, 22, 1996, pp.1-38.

Wu, H., and M. Zhou, "Synonymous Collocation Extraction Using Translation Information," *The 4Jth annual conference of the Association for Computational Linguistics*, 2003, pp. 120-127

Related web pages:

Deja–Vu (http://www.atril.com/).

TOTALrecall (http://candle.cs.nthu.edu.tw/Counter/Counter.asp?funcID=1).

Transit (http://www.star-group.net/eng/software/sprachtech/transit.html).

TransSearch (http://www.tsrali.com/).