Computational Linguistics and Chinese Language Processing
Vol. 9, No. 1 , February 2004, pp. 1-20

Bilingual Collocation Extraction Based on Syntactic and Statistical Analyses

Chien-Cheng Wu[*], and Jason S. Chang[*]

Abstract

In this paper, we describe an algorithm that employs syntactic and statistical analysis to extract bilingual collocations from a parallel corpus. Collocations are pervasive in all types of writing and can be found in phrases, chunks, proper names, idioms, and terminology. Therefore, automatic extraction of monolingual and bilingual collocations is important for many applications, including natural language generation, word sense disambiguation, machine translation, lexicography, and cross language information retrieval.

Collocations can be classified as lexical or grammatical collocations. Lexical collocations exist between content words, while a grammatical collocation exists between a content word and function words or a syntactic structure. In addition, bilingual collocations can be rigid or flexible in both languages. Rigid collocation refers to words in a collocation must appear next to each other, or otherwise (flexible/elastic). We focus in this paper on extracting rigid lexical bilingual collocations. In our method, the preferred syntactic patterns are obtained from idioms and collocations in a machine-readable dictionary. Collocations matching the patterns are extracted from aligned sentences in a parallel corpus. We use a new alignment method based on punctuation statistics for sentence alignment. The punctuation-based approach is found to outperform the length-based approach with precision rates approaching 98%. The obtained collocations are subsequently matched up based on cross-linguistic statistical association. Statistical association between the whole collocations as well as words in collocations is used to link a collocation with its counterpart collocation in the other language. We implemented the proposed method on a very large Chinese-English parallel corpus and obtained satisfactory results.

[*] Department of Computer Science, National Tsing Hua University
Address: 101, Kuangfu Road, Hsinchu, Taiwan
E-mail: g904374@oz.nthu.edu.tw; jschang@cs.nthu.edu.tw

1. Introduction

Collocations, like terminology, tends to be lexicalized and to have a somewhat more restricted meaning than the surface forms suggest [Justeson and Katz, 1995]. Collocations are recurrent combinations of words that co-occur more often than they normally would based on chance. The words in a collocation may appear next to each other (rigid collocations) or in other locations (flexible/elastic collocations). On the other hand, collocations can also be classified as lexical or grammatical collocations [Benson, Benson, Ilson, 1986]. Lexical collocations exist between content words, while a grammatical collocation exists between a content word and function words or a syntactic structure. Collocations are pervasive in all types of writing and can be found in phrases, chunks, proper names, idioms, and terminology. Collocations in one language are usually difficult to translate directly into another language word for word; therefore, they present a challenge for machine translation systems and second language learners alike.

Automatic extraction of monolingual and bilingual collocations is important for many applications, including natural language generation, word sense disambiguation, machine translation, lexicography, and cross language information retrieval. Hank and Church [1990] pointed out the usefulness of mutual information for identifying monolingual collocations in lexicography. Justeson and Katz [1995] proposed to identify technical terminology based on preferred linguistic patterns and discourse properties of repetition. Among the many general methods presented by Manning and Schutze [1999], the best results can be achieved through filtering based on both linguistic and statistical constraints. Smadja [1993] presented a method called EXTRACT, based on the mean variance of the distance between two collocates , that is capable of computing elastic collocations. Kupiec [1993] proposed to extract bilingual noun phrases using statistical analysis of the co-occurrence of phrases. Smadja, McKeown, and Hatzivassiloglou [1996] extended the EXTRACT approach to handle bilingual collocation based mainly on the statistical measures of the Dice coefficient. Dunning [1993] pointed out the weakness of mutual information and showed that log likelihood ratios are more effective in identifying monolingual collocations, especially when the occurrence count is very low.

Both Smadja and Kupiec used the statistical association between whole collocations in two languages without examining the constituent words. For a collocation and its non-compositional translation equivalent, this approach is reasonable. For instance, with the bilingual collocation (“擠破頭”, “stop at nothing”) shown in Example 1, it will not be helpful to examine the statistical association between “stopping” and “擠” [ji, squeeze] (or “破” [bo, broken] and “頭” [tou, head] for that matter). However, for the bilingual collocation (“減薪”, “pay cut”) shown in Example 2, considering the statistical association between “pay” and “薪” [xin, wage] as well as between “cut” and “減” [jian, reduce] certainly makes sense. Moreover, we have more data with which to make statistical inferences between words than between phrases. Therefore, measuring the statistical association of collocations based on constituent words will help us cope with the data sparseness problem. We will be able to extract bilingual collocations with high reliability even when they appear together in aligned sentences only once or twice.

Example 1

They are stopping at nothing to get their kids into "star schools"

他們擠破頭也要把孩子送進明星小學

Source: 1995/02 No Longer Just an Academic Question: Educational Alternatives Come to Taiwan

Example 2

Not only haven't there been layoffs or pay cuts, the year-end bonus and the performance review bonuses will go out as usual .

不但不虞裁員、減薪，年終獎金、考績獎金還都照發不誤

Source: 1991/01 Filling the Iron Rice Bowl

Since collocations can be rigid or flexible in both languages, there are, in general, three types of bilingual collocation matches. In Example 1, (“擠破頭”, “stop at nothing”) is a pair of rigid collocation, and (“把…送進”, “get … into”) is a pair of elastic collocation. In Example 3 ,(“走…的路線’, “take the path of”) is an example of a pair of elastic and rigid collocations.

Example 3

Lin Ku-fang, a worker in ethnomusicology, worries too, but his way is not to take the path of revolutionizing Chinese music or making it more "symphonic"; rather, he goes directly into the tradition, looking into it for "good music" that has lasted undiminished for a hundred generations.

民族音樂工作者林谷芳也非不感到憂心，但他的方法是：不走國樂改革或「交響化」的路，而是直接面對傳統、從中尋找歷百代不衰的「好聽音樂」。

Source: 1997/05 A Contemporary Connoisseur of the Classical Age--Lin Ku-fang's Canon of Chinese Classical Music

In this paper, we describe an algorithm that employs syntactic and statistical analyses to extract rigid lexical bilingual collocations from a parallel corpus. Here, we focus on bilingual collocations, which have some lexical correlation between them and are rigid in both languages. To cope with the data sparseness problem, we use the statistical association between two collocations as well as that between their constituent words. In Section 2, we describe how we obtain the preferred syntactic patterns from collocations and idioms in a machine-readable dictionary. Examples will be given to show how collocations matching the patterns are extracted and aligned for given aligned sentence pairs in a parallel corpus. We implemented the proposed method in an experiment on the Chinese-English parallel corpus of Sinorama Magazine and obtained satisfactory results. We describe the experiments and our evaluation in section 3. The limitations of the study and related issues are taken up in section 4. We conclude and give future directions of research in section 5.

2. Extraction of Bilingual Collocations

In this chapter, we will describe how we obtain bilingual collocations by using preferred syntactic patterns and associative information. Consider a pair of aligned sentences in a parallel corpus such as that shown in Example 4 below:

Example 4

The civil service rice bowl, about which people always said "you can't get filled up, but you won't starve to death either," is getting a new look with the economic downturn. Not only haven't there been layoffs or pay cuts, the year-end bonus and the performance review bonuses will go out as usual, drawing people to compete for their own "iron rice bowl."

以往一向被認爲「吃不飽、餓不死」的公家飯，值此經濟景氣低迷之際，不但不虞裁員、減薪，年終獎金、考績獎金還都照發不誤，因而促使不少人回頭競逐這隻「鐵飯碗」。

Source: 1991/01 Filling the Iron Rice Bowl

We can extract the following collocations and translation counterparts:

(civil service rice bowl, 公家飯)
(get filled up, 吃…飽)
(starve to death, 餓…死)
(economic downturn, 經濟景氣低迷)
(pay cuts, 減薪)

(year-end bonus, 年終獎金)
(performance review bonuses, 考績獎金)
(iron rice bowl, 鐵飯碗)

In section 2.1, we will first show how that process is carried out for Example 4 using the proposed approach. A formal description of our method will be given in section 2.2.

2.1 An Example of Extracting Bilingual Collocations

To extract bilingual collocations, we first run part of speech tagger on both sentences. For instance, for Example 4, we get the results of tagging shown in Examples 4A and 4B.

In the tagged English sentence, we identify phrases that follow a syntactic pattern from a set of training data of collocations. For instance, “jj nn” is one of the preferred syntactic structures. Thus, ”civil service,” “economic downturn,” “own iron” etc are matched. See Table 1 for more details. For Example 4, the phrases shown in Examples 4C and 4D are considered to be potential candidates for collocations because they match at least two distinct collocations listed in LDOCE:

Example 4A

The/at civil/jj service/nn rice/nn bowl/nn ,/, about/in which/wdt people/nns always/rb said/vbd "/`` you/ppss can/md 't/* get/vb filled/vbn up/rp ,/, but/cc you/ppss will/md 't/* starve/vb to/in death/nn either/cc ,/rb "/" is/bez getting/vbg a/at new/jj look/nn with/in the/at economic/jj downturn/nn ./. Not/nn only/rb have/hv 't/* there/rb been/ben layoffs/nns or/cc pay/vb cuts/nns ,/, the/at year/nn -/in end/nn bonus/nn and/cc the/at performance/nn review/nn bonuses/nn will/md go/vb out/rp as/ql usual/jj ,/, drawing/vbg people/nns to/to compete/vb for/in their/pp$ own/jj "/`` iron/nn rice/nn bowl/nn ./. "/"

Example 4B

以往/Nd 一向/Dd 被/P02 認為/VE2 「/PU 吃/VC 不/Dc 飽/VH 、/PU 餓不死/VR 」/PU 的/D5 公家/Nc 飯/Na ，/PU 值此/Ne 經濟/Na 景氣/Na 低迷/VH 之際/NG ，/PU 不但/Cb 不虞/VK 裁員/VC 、/PU 減薪/VB ，/PU 年終獎金/Na 、/PU 考績/Na 獎金/Na 還都/Db 照/VC 發/VD 不誤/VH ，/PU 因而/Cb 促使/VL 不少/Ne 人/Na 回頭/VA 競逐/VC 這/Ne 隻/Nf 「/PU 鐵飯碗/Na 」/PU

Example 4C

“civil service,” ”rice bowl,” ”iron rice bow,” ”fill up,” ”economic downturn,” ”end bonus,” ”year - end bonus,” ”go out,” ”performance review," " performance review bonus," " pay cut," " starve to death," " civil service rice," " service rice," " service rice bowl," " people always," " get fill," " people to compete," " layoff or pay," " new look," " draw people"

Example 4D

"吃不飽," "餓不死," "公家飯," "經濟景氣," "景氣低迷," "經濟景氣低迷," "裁員," "減薪," "年終獎金," "考績獎金," "競逐," " 鐵飯碗."

Although "new look" and "draw people" are legitimate phrases, they are more like "free combinations" than collocations. That is reflected by their low log likelihood ratio values. For this research, we proceed to determine how tightly the two words in overlapping bigrams within a collocation are associated with each other; we calculate the minimum of the log likelihood ratio values for all the bigrams. Then, we filter out the candidates whose POS patterns appear only once or have minimal log likelihood ratios of less than 7.88. See Tables 1 and 2 for more details.

In the tagged Chinese sentence, we basically proceed in the same way to identify the candidates of collocations, based on the preferred linguistic patterns of the Chinese translations of collocations in an English-Chinese MRD. However, since there is no space delimiter between words, it is at times difficult to say whether a translation is a multi-word collocation or a single word, in which case it should not be considered as a collocation. For this reason, we take multiword and singleton phrases (with two or more characters) into consideration. For instance, in tagged Example 4, we extract and consider these candidates shown in Tables 1 and 2 as the counterparts of English collocations.

Notes that at this point, we have not pinned collocations down but allow overlapping and conflicting candidates such as "經濟景氣," "景氣低迷," and "經濟景氣低迷." See Tables 3 and 4 for more details.

Table 1. The initial candidates extracted based on preferred patterns trained on collocations listed in LDOCE (LDOCE example: the example for the POS pattern in LDOCE; Pattern Count: the number of POS patterns occurring in LDOCE ; Min LLR : the minimal LLR value of every two words in the candidate pairs.)

E-collocation Candidate Pairs	Part of Speech	LDOCE example	Pattern Count	Min LLR
civil service	jj nn	hard cash	1562	496.156856
rice bowl	nn nn	beef steak	1860	99.2231161

iron rice bowl	nn nn nn	tin pan alley	8	66.3654678
filled up	vbn rp	set down	84	55.2837871
economic downturn	jj nn	hard cash	1562	51.8600979
*end bonus	nn nn	beef steak	1860	15.9977283
year - end bonus	nn nn nn	tin pan alley	12	15.9977283
go out	vb rp	bang out	1790	14.6464925
performance review	nn nn	beef steak	1860	13.5716459
performance review bonus	nn nn nn	tin pan alley	8	13.5716459
pay cut	vb nn	take action	313	8.53341082
starve to death	vb to nn	bring to bay	26	7.93262494
civil service rice	jj nn nn	high water mark	19	7.88517791
*service rice	nn nn	beef steak	1860	7.88517791
*service rice bowl	nn nn nn	tin pan alley	8	7.88517791
* people always	nn rb	hand back	24	3.68739176
get filled	vb vbn	stay put	3	1.97585732
* people to compete	nn to vb	order to view	2	1.29927068
* layoff or pay	nn cc vb	wine and dine	14	0.93399125
* new look	jj nn	hard cash	1562	0.63715518
* draw people	vbg nn	dying wish	377	0.03947748

* indicates invalid candidate (with human judgment)

Table 2. The candidates of English collocations based on both preferred linguistic patterns and log likelihood ratios.

E-collocation Candidate Pairs	Part of Speech	LDOCE example	Pattern Count	Min LLR
civil service	jj nn	hard cash	1562	496.156856
rice bowl	nn nn	beef steak	1860	99.2231161
iron rice bowl	nn nn nn	tin pan alley	8	66.3654678
filled up	vbn rp	set down	84	55.2837871
economic downturn	jj nn	hard cash	1562	51.8600979
*end bonus	nn nn	beef steak	1860	15.9977283
year - end bonus	nn nn nn	tin pan alley	12	15.9977283
go out	vb rp	bang out	1790	14.6464925
performance review	nn nn	beef steak	1860	13.5716459
performance review bonus	nn nn nn	tin pan alley	8	13.5716459
pay cut	vb nn	take action	313	8.53341082
starve to death	vb to nn	bring to bay	26	7.93262494
civil service rice	jj nn nn	high water mark	19	7.88517791
*service rice	nn nn	beef steak	1860	7.88517791

*service rice bowl	nn nn nn	tin pan alley	8	7.88517791

* indicates an invalid candidate(based on human judgment)

Table 3. The initial candidates extracted by the Chinese collocation recognizer.

C-collocation Candidate Pairs	POS	LDOCE example	Patter Count	Min LLR
不少 人	Ed Na	本國語	2	550.904793
*被 認爲	PP VE	待考慮	6	246.823964
景氣 低迷	Na VH	視力不良	97	79.8159904
經濟 景氣 低迷	Na Na VH	宗教信仰自由	3	47.2912274
經濟 景氣	Na Na	生活津貼	429	47.2912274
公家 飯	Nc Na	全國大選	63	42.6614685
*不 飽	Dc VH	毫無困難	24	37.3489687
考績 獎金	Na Na	生活津貼	429	36.8090448
不虞 裁員	VJ VA	引起爭吵	3	17.568518
回頭 競逐	VA VC	豎耳傾聽	26	14.7120606
*還都 照	Db VC	無法參與	18	14.1291893
*發 不誤	VD VH	供應充份	2	13.8418648
*低迷 之際	VH NG	兩可之間	10	11.9225789
*值此 經濟 景氣	VA Na Na	浮球活栓	2	9.01342071
*值此 經濟	VA Na	劃線支票	94	9.01342071
*照 發	VC VD	登記歸還	2	6.12848087
*人 回頭	Na VA	安危未卜	27	1.89617179

* indicates an invalid candidate (based on human judgment)

Table 4. The result of Chinese collocation candidates which are picked out. (The ones which have no Min LLR are singleton phrases.)

C-collocation Candidate Pairs	POS	LDOCE example	Patter Count	Min LLR
不少 人	Ed Na	本國語	2	550.904793
*被 認爲	PP VE	待考慮	6	246.823964
景氣 低迷	Na VH	視力不良	97	79.8159904
經濟 景氣 低迷	Na Na VH	宗教信仰自由	3	47.2912274
經濟 景氣	Na Na	生活津貼	429	47.2912274
公家 飯	Nc Na	全國大選	63	42.6614685
*不 飽	Dc VH	毫無困難	24	37.3489687
考績 獎金	Na Na	生活津貼	429	36.8090448
不虞 裁員	VJ VA	引起爭吵	3	17.568518

回頭 競逐	VA VC	豎耳傾聽	26	14.7120606
*還都 照	Db VC	無法參與	18	14.1291893
*發 不誤	VD VH	供應充份	2	13.8418648
*低迷 之際	VH NG	兩可之間	10	11.9225789
*值此 經濟 景氣	VA Na Na	浮球活栓	2	9.01342071
*值此 經濟	VA Na	劃線支票	94	9.01342071
之際	NG		5	
經濟	Na		1408	
景氣	Na		1408	
年終獎金	Na		1408	
考績	Na		1408	
獎金	Na		1408	
鐵飯碗	Na		1408	
公家	Nc		173	
以往	Nd		48	
值此	VA		529	
裁員	VA		529	
回頭	VA		529	
減薪	VB		78	
競逐	VC		1070	
認為	VE		139	
低迷	VH		731	
不誤	VH		731	
不虞	VJ		205	
促使	VL		22	
餓不死	VR		14	

To align collocations in both languages, we employ the Competitive Linking Algorithm proposed by Melamed [1996] to conduct word alignment. Basically, the proposed algorithm **CLASS**, the Collocation Linking Algorithm based on Syntax and Statistics, is a greedy method that selects collocation pairs. The pair with the highest association value takes precedence over those with lower values. CLASS also imposes a one-to-one constraint on the collocation pairs selected. Therefore, the algorithm at each step considers only pairs with words that haven't been selected previously. However, CLASS differs with CLA(Competitive Linking Algorithm) in that it considers the association between the two candidate collocations based on two measures:

- the Logarithmic Likelihood Ratio between the two collocations in question as a whole;
- the translation probability of collocation based on constituent words.

In the case of Example 4, the CLASS algorithm first calculates the counts of collocation candidates in the English and Chinese parts of the corpus. The collocations are matched up randomly across from English to Chinese. Subsequently, the co-occurrence counts of these candidates matched across from English to Chinese are also tallied. From the monolingual collocation candidate counts and cross language concurrence counts, we produce the LLR values and the collocation translation probability derived from word alignment analysis. Those collocation pairs with zero translation probability are ignored. The lists are sorted in descending order of LLR values, and the pairs with low LLR value are discarded. Again, in the case of Example 4, the greedy selection process of collocation starts with the first entry in the sorted list and proceeds as follows:

1. The first, third, and fourth pairs, ("iron rice bowl," "鐵飯碗"), ("year-end bonus," "年終獎金"), and ("economic downturn," "經濟景氣低迷"), are selected first. Thus, conflicting pairs will be excluded from consideration, including the second pair, fifth pair and so on.
2. The second entry ("rice bowl," "鐵飯碗"), fifth entry ("economic downturn," "值此經濟景氣") and so on conflict with the second and third entries that have already been selected. Therefore, CLASS skips over these entries.
3. The entries ("performance review bonus," "考績獎金"), ("civil service rice," "公家飯"), ("pay cuts," "減薪"), and ("starve to death," "餓不死") are selected next.
4. CLASS proceeds through the rest of the list and the other list without finding any entries that do not conflict with the seven entries previously selected.
5. The program terminates and outputs a list of seven collocations.

Table 5. The extracted Chinese collocation candidates which are picked out. The shaded collocation pairs are selected by CLASS (Greedy Alignment Linking E).

English collocations	Chinese collocations	LLR	Collocation Translation Prob.
iron rice bowl	鐵飯碗	103.3	0.0202
rice bowl	鐵飯碗	77.74	0.0384
year-end bonus	年終獎金	59.21	0.0700
economic downturn	經濟 景氣 低迷	32.4	0.9359
economic downturn	值此 經濟 景氣	32.4	0.4359
...	. . .	...	...
performance review bonus	考績 獎金	30.32	0.1374
economic downturn	景氣 低迷	29.82	0.2500
civil service rice	公家 飯	29.08	0.0378

pay cuts	減薪	28.4	0.0585
year-end bonus	考績 獎金	27.35	0.2037
performance review	考績	27.32	0.0039
performance review bonus	年終獎金	26.31	0.0370
starve to death	餓不死	26.31	0.5670
...	. . .	...	...
rice bowl	公家 飯	24.98	0.0625
iron rice bowl	公家 飯	25.60	0.0416
…	…	…	…

2.2 The Method

In this section, we describe formally how CLASS works. We assume the availability of a parallel corpus and a list of collocations in a bilingual MRD. We also assume that the sentences and words have been aligned in the parallel corpus. We will describe how **CLASS** extracts bilingual collocations from such a parallel corpus. CLASS carries out a number of preprocessing steps to calculate the following information:

1. lists of preferred POS patterns of collocation in both languages;
2. collocation candidates matching the preferred POS patterns;
3. n-gram statistics for both languages, N = 1, 2;
4. log likelihood ratio statistics for two consecutive words in both languages;
5. log likelihood ratio statistics for a pair of candidates of bilingual collocations across one language to the other;
6. content word alignment based on the Competitive Linking Algorithm [Melamed, 1997.]

Figure 1 illustrates how the method works for each aligned sentence pair (*C*, *E*) in the corpus. Initially, part of speech taggers process *C* and *E*. After that, collocation candidates are extracted based on preferred POS patterns and statistical association between consecutive words in a collocation. The collocation candidates are subsequently matched up from one language to the other. These pairs are sorted according to the log likelihood ratio and collocation translation probability. A greedy selection process goes through the sorted list and selects bilingual collocations subject to one-to-one constraint. The detailed algorithm is given below:

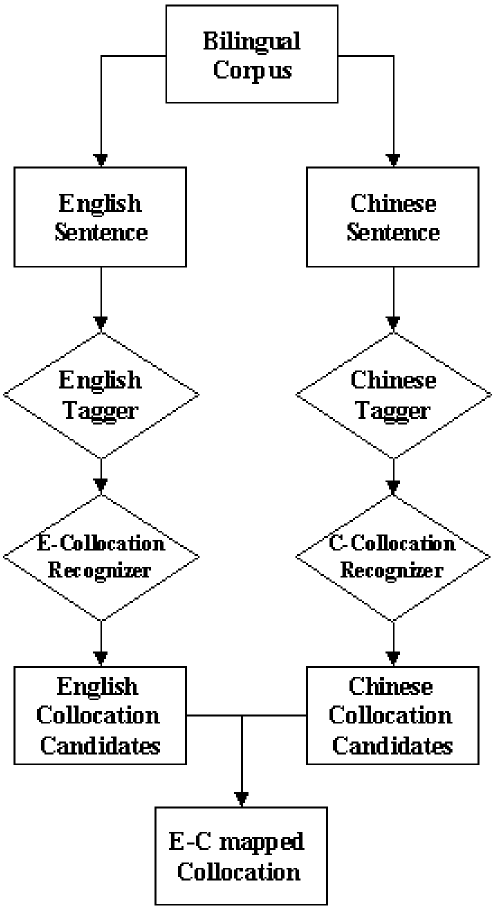

Figure 1. The major components in the proposed CLASS algorithm.

Preprocessing: Extracting preferred POS patterns *P* and *Q* in both languages

Input: A list of bilingual collocations from a machine-readable dictionary

Output:

1. Perform part of speech tagging for both languages.
2. Calculate the number of instances for all POS patterns in both languages.
3. Eliminate the POS patterns with instance counts of 1.

Collocation Linking Alignment based on Syntax and Statistics

Extract bilingual collocations from aligned sentences.

Input:

(1) A pair of aligned sentences (C, E), $C = (C_1\ C_2 \dots C_n)$ and $E = (E_1\ E_2 \dots E_m)$.

(2) Preferred POS patterns P and Q in both languages.

Output: Aligned bilingual collocations in (C, E)

1. C is segmented and tagged with part of speech information T.
2. E is tagged with part of speech sequences S.
3. Match T against P and match S against Q to extract collocation candidates $X_1, X_2, \dots X_k$ in English and $Y_1, Y_2, \dots, Y_e$ in Chinese.
4. Consider each bilingual collocation candidate (X_i, Y_j) in turn and calculate the minimal log likelihood ratio LLR between X_i and Y_j:

   MLLR (D) = $\min_{i=1,n-1} LLR(W_i, W_{i+1})$
5. Eliminate candidates with LLR that are smaller than a threshold (7.88).
6. Match up all possible links from English collocation candidates to Chinese ones: (D_1, F_1), (D_1, F_2), … (D_i, F_j), … (D_m, F_n).
7. Calculate LLR for (D_i, F_j) and discard pairs with LLR value that are lower than 7.88.
8. The only candidate list of bilingual collocations considered is the one with non-zero collocation translation probability $P(D_i, F_j)$ values. The list is then sorted based on the LLR values and collocation translation probability.
9. Go down the list and select a bilingual collocation if it does not conflict with a previous selection.
10. Output the bilingual collocation selected in Step 9.

Log-likelihood ratio: LLR(x;y)

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

k_1 : # of pairs that contain x and y simultaneously.
k_2 : # of pairs that contain x but do not contain y.
n_1 : # of pairs that contain y
n_2 : # of pairs that does not contain y
$p_1 =$ k_1/n_1, $p_2 = k_2/n_2$,
$p =$ $(k_1+k_2)/(n_1+n_2)$

Collocation translation probability P(x | y)

$$P(D_i \mid F_j) = \frac{1}{k}\sum_{e \in F_j} \max_{c \in D_i} P(c \mid e)$$

k : number of words in the English collocation F_j

3. Experiments and Evaluation

We have implemented CLASS using the Longman Dictionary of Contemporary English, English-Chinese Edition, and the parallel corpus of Sinorama magazine. The articles from

Sinorama covered a wide range of topics, reflecting the personalities, places, and events in Taiwan for the previous three decades. We experimented on articles mainly dating from 1995 to 2002. Sentence and word alignment were carried out first to obtain the Sinorama Parallel Corpus.

Sentence alignment is a very important aspect of CLASS. It is the basis for good collocation alignment. We use a new alignment method based on punctuation statistics [Yeh & Chang, 2002]. The punctuation-based approach has been found to outperform the length-based approach with precision rates approaching 98%. With the sentence alignment approach, we obtained approximately 50,000 reliably aligned sentences containing 1,756,000 Chinese words (about 2,534,000 Chinese characters) and 2,420,000 English words in total.

The content words were aligned using the Competitive Linking Algorithm. Alignment of content words resulted in a probabilistic dictionary with 229,000 entries. We evaluated 100 random sentence samples with 926 linking types, and the achieved precision rate was 93.3%. Most of the errors occurred with English words having no counterpart in the corresponding Chinese sentence. Translators do not always translate word for word. For instance, with the word “water” in Example 5, it seems that there is no corresponding pattern in the Chinese sentence. Another major cause of errors was collocations that were not translated compositionally. For instance, the word “State” in the Example 6 is a part of the collocation “United States,” and “美國” is more highly associated with “United” than “States”; therefore, due to the one-to-one constraint “States” will not be aligned with “美國”. Most often, it will be aligned incorrectly. About 49% of the error links were of this type.

Example 5

The boat is indeed a vessel from the mainland that illegally entered Taiwan waters. The words were a "mark" added by the Taiwan Garrison Command before sending it back.

編按：此船的確是大陸偷渡來台船隻，那八個字只不過是警總在遣返前給它加的「記號」！

Source: 1990/10 Letters to the Editor

Example 6

Figures issued by the American Immigration Bureau show that most Chinese immigrants had set off from Kwangtung and Hong Kong, which is why the majority of overseas Chinese in the United States to this day are of Cantonese origin.

由美國移民局發表的數字來看，中國移民以從廣東、香港出海者最多，故到現在爲止，美國華僑仍以原籍廣東者佔大多數。

Source: 1990/09 All Across the World: The Chinese Global Village

We obtained the word-to-word translation probability from the result of word alignment. The translation probability P($c|e$) is calculated as followed:

$$P(c|e) = \frac{count\ (e,c)}{count\ (e)}, \text{ where}$$

count(e,c) : the number of alignment links between a Chinese word *c* and an English word *e*;

count(e) : the number of instances of e in alignment links.

Take “pay” as an example. Table 6 shows the various alignment translations for “pay” and the translation probability.

Table 6. The aligned translations for the English word “pay” and their translation probability.

Translation	Count	Translation Prob.	Translation	Count	Translation Prob.
代價	34	0.1214	花錢	7	0.025
錢	31	0.1107	出錢	6	0.0214
費用	21	0.075	租	6	0.0214
付費	16	0.0571	發給	6	0.0214
領	16	0.0571	付出	5	0.0179
繳	16	0.0571	薪資	5	0.0179
支付	13	0.0464	付錢	4	0.0143
給	13	0.0464	加薪	4	0.0143
薪水	11	0.0393	. . .	...	...
負擔	9	0.0321	積欠	2	0.0071
費	9	0.0321	繳款	2	0.0071
給付	8	0.0286			

Before running CLASS, we obtained 10,290 English idioms, collocations, and phrases together with 14,945 Chinese translations in LDOCE. After part of speech taggi ng, we had 1,851 distinct English patterns and 4326 Chinese patterns. To calculate the statistical association within words in a monolingual collocation and across the bilingual collocations, we built N-grams for the Sinorama Parallel Corpus. There were 790,000 Chinese word bigrams and 669,000 distinct English bigrams. CLASS identified around 595,000 Chinese collocation candidates (184,000 distinct types) and 230,000 English collocation candidates (135,000 distinct types) through this process.

We selected 100 sentences to evaluate the performance. We focused on rigid lexical collocations. The average English sentence had 45.3 words, while the average Chinese sentence had 21.4 words. The two human judges, both master students majoring in Foreign Languages, identified the bilingual collocations in these sentences. We then compared the bilingual collocations produced by CLASS against the answer keys. The evaluation produced an average recall rate = 60.9 % and precision rate = 85.2 % (see Table 7).

Table 7. Experiment results of bilingual collocation from the Sinorama Parallel Corpus.

# keys	#answers	#hits	#errors	Recall	Precision
382	273	233	40	60.9%	85.2%

4. Discussion

This paper describes a new approach to the automatic acquisition of bilingual collocations from a parallel corpus. Our method is an extension of Melamed’s Competitive Linking Algorithm for word alignment. It combines both linguistic and statistical information and uses it to recognize monolingual and bilingual collocations in a much simpler way than Smadja’s work does. Our approach differs from previous work in the following ways:

1. We use a data-driven approach to extract monolingual collocations.
2. Unlike Smadja and Kupiec, we do not commit to two sets of monolingual collocations. Instead, we consider many overlapping and conflicting candidates and rely on cross linguistic statistics to revolve the issue.
3. We combine both type of information related to the whole collocation as well as to the constituent words to achieve more reliable probabilistic estimation of aligned collocations.

Our approach is limited by its reliance on training data consisting of mostly rigid collocation patterns, and it is not applicable to elastic collocations such as “jump on … bandwagon.” For instance, the program cannot handle the elastic collocation in the following example:

Example 7

台灣幸而趕**搭**了一程獲利豐厚的**順風車**，可以將目前剛要起步的馬來西亞、中國大陸等國家遠拋身後。

Taiwan has had the good fortune to **jump on** this high-profit **bandwagon** and has been able to snatch a substantial lead over countries like Malaysia and mainland China, which have just started in this industry.

Source: Sinorama, 1996, Dec Issue Page 22, Stormy Waters for Taiwan’s ICs

This limitation can be partially alleviated by matching nonconsecutive word sequences against existing lists of collocations for the two languages.

Another limitation has to do with bilingual collocations, which are not literal translations. For instance, “difficult and intractable” can not yet be handled by the program, because it is not a word for word translation of “桀傲不馴”.

Example 8

意思是說一個再怎麼**桀傲不馴**的人，都會有人有辦法制服他。

This saying means that no matter how difficult and intractable a person may seem, there will always be someone else who can cut him down to size.

Source: 1990/05 A Fierce Horse Ridden by a Fierce Rider

In the experiment, we found that this limitation may be partially solved by splitting the candidate list of bilingual collocations into two lists: one (NZ) with non-zero phrase translation probabilistic values and the other (ZE) with zero values. The two lists can then be sorted based on the LLR values. After extracting bilingual collocations from the NZ list, we could continue to go down the ZE list and select bilingual collocations that did not conflict with previously selection.

In the proposed method, we do no take advantage of the correspondence between POS patterns in one language with those in the other. Some linking mistakes seem to be avoidable if POS information is used. For example, the aligned collocation for “issue/**vb** visas/**nns**” is “簽證/**Na**”, not “發/**VD** 簽證/**Na.**” However, the POS pattern “vb nn” appears to be more compatible with “VD Na” than with “Na.”

Example 9

一九七二年澳洲承認中共，中華民國即於此時與澳斷交。因爲無正式邦交，澳洲不能在台灣**發簽證**，而由澳洲駐香港的使館代辦，然後將簽證送回台灣，簽證手續約需五天至一周。

The Republic of China broke relations with Australia in 1972, after the country recognized the Chinese Communists, and because of the lack of formal diplomatic relations, Australia felt it could not **issue visas** on Taiwan. Instead, they were handled through its consulate in Hong Kong and then sent back to Taiwan, the entire process requiring five days to a week to complete.

Source: 1990/04 Visas for Australia to Be Processed in Just 24 Hours

A number of mistakes are caused by erroneous word segments in the Chinese tagger. For instance, "大學及研究生出國期間" should be segmented as " 大學 / 及 / 研究生 / 出國 / 期間" but instead is segmented as "大學 / 及 / 研究 / 生出 / 國 / 期間 / 的 / 學業." Another major source of segmentation mistakes has to do with proper names and their transliterations. These name entities that are not included in the database are usually segmented into single Chinese characters. For instance, "...一書作者劉學銚指出..." is segmented as " ... / 一 / 書 / 作者 / 劉 / 學 / 銚 / 指出 / ...," while "...在匈牙利地區建國的馬札爾人..." is segmented as "...在 / 匈牙利 / 地區 / 建國 / 的 / 馬 / 札 / 爾 / 人 / ...." Therefore, handling these name entities in a pre-process should be helpful to avoid segmenting mistakes and alignment difficulties.

5. Conclusion and Future Work

In this paper, we have presented an algorithm that employs syntactic and statistical analyses to extract rigid bilingual collocations from a parallel corpus. Phrases matching the preferred patterns are extracted from aligned sentences in a parallel corpus. These phrases are subsequently matched up based on cross-linguistic statistical association. Statistical association between the whole collocations as well as words in the collocations is used jointly to link a collocation with its counterpart. We implemented the proposed method on a very large Chinese-English parallel corpus and obtained satisfactory results.

A number of interesting future directions suggest themselves. First, it would be interesting to see how effectively we can extend the method to longer and elastic collocations and to grammatical collocations. Second, bilingual collocations that are proper names and transliterations may need additional consideration. Third, it will be interesting to see if the performance can be improved using cross language correspondence between POS patterns.

References

Benson, Morton., Evelyn Benson, and Robert Ilson." The BBI Combinatory Dictionary of English: A Guide to Word Combinations. " *John Benjamins, Amsterdam, Netherlands,* 1986.

Choueka, Y. "Looking for needles in a haystack", *RIAO, Conference on User-Oriented Context Based Text and Image Handling, Cambridge*, 1988, pp. 609-623.

Choueka, Y.; Klein, and Neuwitz, E.. "Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus." *Journal of the Association for Literary and Linguistic Computing*, 4(1), 1983, pp34-38.

Church, K. W. and Hanks, P. "Word association norms, mutual information, and lexicography." *Computational Linguistics*, 16(1) , 1990, pp. 22-29.

Dagan, I. and K. Church. "Termight: Identifying and translation technical terminology". In *Proc. of the 4th Conference on Applied Natural Language Processing (ANLP)*, 1994, pages 34-40.

Dunning, T. "Accurate methods for the statistics of surprise and coincidence", *Computational Linguistics* 19:1, 1993, pp.61-75.

Haruno, M., S. Ikehara, and T. Yamazaki. "Learning bilingual collocations by word-level sorting." *In Proc. of the 16th International Conference on Computational Linguistics (COLING '96)*, 1996, pp. 525-530.

Huang, C.-R., K.-J. Chen, Y.-Y. Yang, "Character-based Collocation for Mandarin Chinese", *In Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000, pp. 540-543.

Inkpen, Diana Zaiu and Hirst, Graeme. "Acquiring collocations for lexical choice between near-synonyms." *In Proceedings of the Workshop on Unsupervised Lexical Acquisition, 40th Annual Meeting of the Association for Computational Lin- guistics (ACL 2002)*, 2002, pp. 67-76.

Justeson, J.S. and Slava M. Katz. "Technical Terminology: some linguistic properties and an algorithm for identification in text." *Natural Language Engineering*, 1(1), 1995, pp. 9-27.

Kupiec, Julian. "An algorithm for finding noun phrase correspondences in bilingual corpora." *In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 1993, pp. 17-22.

Lin, D. "Using collocation statistics in information extraction." *In Proc. of the Seventh Message Understanding Conference (MUC-7)*, 1998.

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

Melamed, I. Dan. "A Word-to-Word Model of Translational Equivalence." *In Proceedings of the 35st Annual Meeting of the Association for Computational Linguistics*, 1997, pp 490-497.

Smadja, F. "Retrieving collocations from text: Xtract." *Computational Linguistics*, 19(1) 1993, pp143-177.

Smadja, F., K.R. McKeown, and V. Hatzivassiloglou. "Translating collocations for bilingual lexicons: A statistical approach." *Computational Linguistics*, 22(1) ,1996, pp 1-38.

Yeh, "Using Punctuation Marks for Bilingual Sentence Alignment." Master thesis, 2003, National Tsing Hua University, Taiwan