International Journal of Computer Processing of Oriental Languages
Vol. 20, Nos. 2 & 3 (2007) 1–16
© Chinese Language Computer Society &
World Scientific Publishing Co.

World Scientific
www.worldscientific.com

Building a Collocation Net[a]

ZHOU GUODONG[*,†], ZHANG MIN[†], LI JUNHUI[*] AND ZHU QIAOMING[*]

[*]School of Computer Science and Technology,
Soochow University, 1 Shizi Street, Suzhou, China 215006
{gdzhou, jhli, qmzhu}@suda.edu.cn
[†]Institute for Infocomm Research,
21 Heng Mui Keng Terrace, Singapore 119613
{zhougd, mzhang}@i2r.a-star.edu.sg

This paper presents an approach to build a novel two-level collocation net, which enables calculation of the collocation relationship between any two words, from a large raw corpus. The first level consists of atomic classes (each atomic class consists of one word and feature bigram), which are clustered into the second level class set. Each class in both levels is represented by its collocation candidate distribution, extracted from the linguistic analysis of the raw training corpus, over possible collocation relation types. In this way, all the information extracted from the linguistic analysis is kept in the collocation net. Our approach applies to both frequently and less-frequently occurring words by providing a clustering mechanism and resolve the data sparseness problem through the collocation net. Experimentation shows that the collocation net is efficient and effective in solving the data sparseness problem and determining the collocation relationship between any two words.

Keywords: Collocation net; Data sparseness problem; Clustering.

1. Introduction

In any natural language, there always exist many highly associated relationships between words. The two words "strong" and "powerful" are perhaps the canonical example. Although "strong" and "powerful" have similar syntax and semantics, there exist contexts where one is much more appropriate than the other [6].

[a] Part of the work was done when the author was at the Institute for Infocomm Research, Singapore.

For example, we always say “strong tea” instead of “strong computer” and “powerful computer” instead of “powerful tea”. Psychological experiments [11] also indicated that human’s reaction to a highly associated word pair was stronger and faster than that to a poorly associated one. Lexicographers use the terms “collocation” and “co-occurrence” to describe various constraints on pairs of words. Here, we restrict “collocation” in the narrower sense between grammatically bound words, e.g. “strong” and “tea”, which occur in a particular grammatical order, and “co-occurrence” for the more general phenomenon of relationships between words, e.g. “doctor” and “nurse”, which are likely to be used in the same context [10].

This paper will concentrate on “collocation” rather than “co-occurrence” although there is much overlap between these two terms. There is also considerable overlap between the concept of “collocation” and notions like “term”, “technical term” and “terminological phrase”. The latter three are commonly used when collocations are extracted from technical domain. However, it should be noted that the word “term” has a different meaning in information retrieval, where it refers to words and phrases.

There are more and more interest in collocations and co-occurrences partly because this area has been undervalued in the structural linguistic traditions that follow Saussure and Chomsky. Structural linguistics concentrates on the general abstractions about properties of phrases and sentences. In contrast, *Contextual Theory of Meaning* that follows Firth, Halliday and Sinclair, emphasizes the importance of context: the context of social setting, the context of discourse and the context of surrounding words. Such detailed contextual information easily gets lost in structural linguistics. This paper follows Firth’s Contextual Theory of Meaning to discover the collocations, which are grammatically bound. Collocations are important for a number of applications: natural language generation, computational lexicography, parsing, proper noun discovery, corpus linguistic research, machine translation, information retrieval, etc. As an example, [7] showed how syntactic related collocation statistics can be used to improve the performance of the parser on sentences such as “She wanted/placed/put the dress on the rack.”, where lexical preferences are crucial to resolving the ambiguity of prepositional phrase attachment. It also showed that a parser can enforce these preferences by comparing the statistical association of the syntactic relation verb-preposition (“want…on”) with the statistical association of the syntactic relation object-preposition (“dress…on”), when attaching the prepositional phrase.

Currently, there are two categories of approaches used to discover collocations and co-occurrences: statistics-based and parsing-based.

1.1. Statistics-based methods

The methods in this category are normally used to extract the word co-occurrence relationship, the phenomena where words are likely to occur in the same context, from the raw unparsed corpus. Different criteria are used to determine the word co-occurrences. First of all, frequency-based method [12, 8, 18] uses the frequencies of the word pairs with the optional help of part-of-speech filter, stop word list and/or acceptable patterns. Secondly, mean and variance-based method [14] computes the mean and variance of the offsets between the words in the corpus, and the word pairs, which have low variances, are regarded as word co-occurrences. Thirdly, hypothesis testing-based methods are used to determine whether two words occur in the same context more than chance. For example, t-test [1, 3] assumes normal distribution and looks at the difference between the observed and expected means, scaled by the variance of the sample data. Chi-square test [2, 15] uses n-by-n table to show the dependence of occurrences between words and compares the observed frequencies in the table with the expected frequencies for independence. Likelihood ratio [5] assumes binomial distribution and tells how more likely the independence hypothesis is than the dependence hypothesis. Fourthly, mutual information-based method [13, 17, 19, 20] tells the change of information when two words co-occur.

However, there exist several problems with the statistics-based methods:

- These methods cannot differentiate between different types of linguistic relations and the extracted co-occurrences may not be grammatically bound.
- These methods are only effective on frequently occurred words and not effective on less frequently occurred words because they provide no mechanism for categorization to resolve the problem of sparseness.
- The co-occurrences extracted by the frequency-based method and the variance-based method may be negative. The scores used by t-test and chi-square are difficult to interpret while mutual information-based method is the worst for the low frequently occurred words.
- The extracted co-occurrences are always stored in a dictionary, which only contains a limited number of entries and very limited information for each one.

1.2. Parsing-based Methods

The parsing-based methods rely on the syntactic analysis. These methods can extract linguistic related word collocations from the parsed trees and can differentiate between different types of linguistic relations. Normally these methods are combined with the frequency-based method to reject the ones whose frequencies are below the predefined threshold [16].

However, there also exist several problems with the parsing-based methods:

- These methods only apply to frequently occurred words and are not effective on less frequently occurred words because they provide no mechanism for categorization to resolve the problem of data sparseness.
- Manual intervention may be required to ensure that the extracted collocations are valid especially when other statistics-based methods, e.g. the frequency-based method, is not used to reject the invalid ones.
- Similar to the co-occurrences extracted using the statistics-based methods, the collocations extracted using the parsing-based methods may be negative and are always stored in a dictionary, which only contains a limited number of entries and very limited information for each one.

Generally, both the statistics and parsing-based approaches are only effective on frequently occurring words and not effective on less frequently occurring words due to the data sparseness problem. Moreover, the extracted collocations or co-occurrences are always stored in a dictionary, which only contains a limited number of entries with limited information for each one. Finally, the collocation dictionary normally does not differentiate the strength of various collocations.

This paper combines the parsing-based approach and the statistics-based approach, and proposes a novel structure of collocation net. Through the collocation net, the data sparseness problem is resolved by providing a clustering mechanism and the collocation relationship between any two words can be easily determined and measured from the collocation net. Here, the collocation relationship is calculated using novel estimated pair-wise mutual information (EPMI) and estimated average mutual information (EAMI). Moreover, all the information extracted from the linguistic analysis is kept in the collocation net. Compared with the traditional collocation dictionary, the collocation net provides a much more powerful facility since it can determine and measure the collocation relationship between any two words quantitatively.

The layout of this paper is as follows: Section 2 describes the novel structure of collocation net. Section 3 describes estimated pair-wise mutual information (EPMI) and estimated average mutual information (EAMI) to determine and measure the collocation relationship between any two words while Section 4 presents a method for automatically building a collocation net given a large law corpus. Experimentation is given in Section 5. Finally, some conclusions are drawn in Section 6.

2. Collocation Net

The collocation net is a kind of two-level structure, which stores rich information about the collocation candidates and others extracted from the linguistic analysis of a large raw corpus. The first level consists of word and feature bigrams[b] while the second level consists of classes that are clustered from the word and feature bigrams in the first level. For convenience, each word and feature bigram in the first level is also regarded as a class (atomic class). That is to say, each first level atomic class contains only one bigram while each second level class contains one or more word and feature bigrams clustered from first level atomic classes.

Meanwhile, each class in both levels of the collocation net is represented by its related collocation candidate distribution, extracted from the linguistic analysis. In this paper, a collocation candidate is represented as a 3tuple: a left side, a right side and a collocation relation type, which represents the collocation relationship between the left side and the right side. Both the left and right sides can be either a word and feature bigram or a class of word and feature bigrams. For example, a collocation candidate can be either $wf_i - CR_k - wf_j$ or $C_{hi} - CR_k - C_{gi}$, where wf_i is a word and feature bigram; C_{hi} is the ith class in the hth level and CR_k is a relation type.

Briefly, the collocation net is defined as follows:

$$CoNET = \{wf,\ CR,\ L1,\ L2,\ P_{h\text{->}g}\} \tag{1}$$

- wf stores possible word and feature bigrams extracted from the syntactic analysis: $wf = \{wf_i, 1 \le i \le |wf|\}$, where wf_i is the ith word and feature pair in wf and $|wf|$ is the number of the word and feature pairs in wf.
- CR stores possible collocation relation types returned from the syntactic analysis: $CR = \{CR_i, 1 \le i \le |CR|\}$, where CR_i the ith linguistic relation in CR and $|CR|$ is the number of the collocation relation types in CR.
- $L1$ and $L2$ are the first and second levels in the collocation net, respectively;

$$L_h = \{\langle C_{hi}, FDCC_{C_{hi}} \rangle\} \tag{2}$$

where $C_h = \{C_{hi}, 1 \le i \le |C_h|\}$ is the class set in L_h (Obviously, $C_1 = wf$); C_{hi} is the ith class in C_h; $|C_h|$ is the number of the classes in C_h and $FDCC_{C_{hi}} (1 \le i \le |C_h|)$ is the frequency distribution of collocation candidates

[b] The reason to use the word and feature bigram is to distinguish the same word with different features, which can be "word sense", "part-of-speech", etc. In this paper, "part-of-speech" is used as the feature.

related to C_{hi}. That is, each word and feature bigram or class in the collocation net is represented by the distribution of its related collocation candidates. In this way, all the information extracted via the linguistic analysis is stored in the collocation net.

- $P_{h\text{->}g}$ defines the class transition probability from a class to another class in the collocation net:

$$P_{h->g} = \{P(C_{hi} \to C_{gj}), 1 \le i \le |C_h|, 1 \le j \le |C_g|\}, \tag{3}$$

where $P(C_{hi} \to C_{gj})$ is the class transition probability from C_{hi} to C_{gj}. In this paper, $P(C_{hi} \to C_{gj})$ is defined using a kernel function $k(C_{gj}, C_{hi})$[c], which measures the similarity between C_{gj} and C_{hi} by calculating their shared probability distribution of collocation candidates:

$$\begin{aligned} P(C_{hi} \to C_{gj}) &= k(C_{hi}, C_{gj}) \\ &= \sum_{k=1}^{|CR|}\sum_{m=1}^{|wf|} \min(P(C_{hi} - CR_k - wf_m | C_{hi}), P(C_{gj} - CR_k - wf_m | C_{gj})) \\ &\quad + \sum_{k=1}^{|CR|}\sum_{m=1}^{|wf|} \min(P(wf_m - CR_k - C_{hi} | C_{hi}), P(wf_m - CR_k - C_{gj} | C_{gj})). \end{aligned} \tag{4}$$

Through the class transition probability facility $P_{h\text{->}g}$, the collocation net provides a word-clustering mechanism to resolve the data sparseness problem and becomes effective in determining and measuring the collocation relationship between any two words, whether they are frequently occurring or not. We will discuss this in more detail in Section 3.

3. Measuring a Collocation Candidate

For a collocation candidate, e.g. $wf_i - CR_k - wf_j$, we propose two quantitative measurements to calculate the collocation relationship between the two word and feature bigrams wf_i and wf_j given the collocation relation type CR_k: estimated pairwise mutual information (EPMI) and estimated average mutual information (EAMI). Moreover, we also extend the EPMI and EAMI to determine and measure the collocation relationship between any two words. In this way, we can not only determine the most possible collocation relationship between any two words but also measure the strength of the collocation relationship between them.

[c] Other kernel functions may apply. We will study more possibilities in the near future.

3.1. EAMI: Estimated Average Mutual Information

Traditionally in information theory, average mutual information (AMI) measures the co-occurrence relationship between two words as follows:

$$AMI(w_i, w_j) = P(w_i, w_j) \cdot \log \frac{P(w_i, w_j)}{P(w_i) \cdot P(w_j)}. \tag{5}$$

For a collocation candidate, e.g. $wf_i - CR_k - wf_j$, we can extend the above notion and measure the average mutual information between the two word and feature bigrams wf_i and wf_j given the collocation relation type CR_k as follows:

$$\begin{aligned} AMI(wf_i - CR_k - wf_j) &= P(wf_i - CR_k - wf_j \mid CR_k) \\ &\quad \cdot \log \frac{P(wf_i - CR_k - wf_j \mid CR_k)}{P(wf_i - CR_k -^{*} \mid CR_k) \cdot P(^{*} - CR_k - wf_j \mid CR_k)}. \end{aligned} \tag{6}$$

Here, we use "*" to indicate all the possibilities on the corresponding part. The problem with the above equation is that it only works on frequently occurring word and feature bigrams and is not reliable on less-frequently occurring word and feature bigrams (e.g. frequency < 100). In order to resolve this problem, we propose a modified version of AMI, called estimated average mutual information (EAMI), to measure the collocation relationship of a collocation candidate when one or two word and feature bigrams do not occur frequently. This is done by finding two optimal classes in the collocation net and mapping the less-frequently occurring word and feature bigrams to them through the word-clustering mechanism provided in the collocation net as follows:

$$\begin{aligned} &EAMI(wf_i - CR_k - wf_j) = EAMI(C_{1i} - CR_k - C_{1j}) \\ &= \max_{\substack{C_{1i} \to C_{hm}, \\ C_{1j} \to C_{gn}}} \Bigg\{ P(C_{hm} - CR_k - C_{gn} \mid CR_k) \cdot P(C_{1i} \to C_{hm}) \cdot P(C_{1j} \to C_{gn}) \\ &\quad \cdot \log \frac{P(C_{hm} - CR_k - C_{gn} \mid CR_k)}{P(C_{hm} - CR_k -^{*} \mid CR_k) \cdot P(^{*} - CR_k - C_{gn} \mid CR_k)} \Bigg\} \\ &= \max_{\substack{C_{1i} \to C_{hm}, \\ C_{1j} \to C_{gn}}} \{ P(C_{1i} \to C_{hm}) \cdot P(C_{1j} \to C_{gn}) \cdot AMI(C_{hm} - LR_k - C_{gn}) \} \end{aligned} \tag{7}$$

where $P(C_{hm} - CR_k - C_{gn} \mid CR_k) \cdot P(C_{1i} \to C_{hm}) \cdot P(C_{1j} \to C_{gn})$ is the estimated joint probability of the collocation candidate $C_{1i} - CR_k - C_{1j}$ given the class transitions $C_{1i} \rightarrow C_{hm}$ and $C_{1j} \rightarrow C_{gn}$. Here, C_{hm} can be either C_{1i} itself or any class in $L2$ while C_{gn} can be either C_{1j} itself or any class in $L2$. That is, C_{1i}/C_{1j} can be either mapped to itself when the word and feature bigram occurs frequently or mapped to any class in $L2$ when the word and feature bigram does not occur frequently.

3.2. EPMI: Estimated Pairwise Mutual Information

Similarly in information theory, pair-wise mutual information (PMI) measures the change of information between two words as follows:

$$PMI(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i) \cdot P(w_j)} \tag{8}$$

For a collocation candidate, e.g. $wf_i - CR_k - wf_j$, we can also extend the above notion to measure the PMI of the collocation candidate as follows:

$$\begin{aligned} & PMI(wf_i - CR_k - wf_j) \\ & \quad = \log \frac{P(wf_i - CR_k - wf_j \,|CR_k)}{P(wf_i - CR_k -{}^{*} \,|CR_k) \cdot P({}^{*} - CR_k - wf_j \,|CR_k)} \end{aligned} \tag{9}$$

Similarly to AMI, the problem with the above equation is that it only works on frequently occurring word and feature bigrams. In order to resolve this problem, we also propose a modified version of PMI, called estimated pair-wise mutual information (EPMI), to calculate the information change of a collocation candidate when one or two word and feature bigrams do not occur frequently. This is done by using the two optimal classes found in calculating EAMI as follows:

$$\begin{aligned} & \underset{C_{hm},C_{gn}}{EPMI}(wf_i - CR_k - wf_j) = \underset{C_{hm},C_{gn}}{EPMI}(C_{1i} - LR_k - C_{1j}) \\ & \quad = P(C_{1i} \rightarrow C_{hm}) \cdot P(C_{1j} \rightarrow C_{gn}) \\ & \qquad \cdot \log \frac{P(C_{hm} - CR_k - C_{gn} \,|CR_k)}{P(C_{hm} - CR_k -{}^{*} \,|CR_k) \cdot P({}^{*} - CR_k - C_{gn} \,|CR_k)} \\ & \quad = P(C_{1i} \rightarrow C_{hm}) \cdot P(C_{1j} \rightarrow C_{gn}) \cdot PMI(C_{hm} - LR_k - C_{gn}) \end{aligned} \tag{10}$$

Equation (10) measures the pair-wise mutual information using the collocation candidate between the two optimal classes and takes the class transitions $C_{1i} \rightarrow C_{hm}$ and $C_{1j} \rightarrow C_{gn}$ into consideration. In this paper, EAMI is used not only as a quantitative measurement for a collocation candidate but also as a selection criteria to determine the two optimal classes in calculating EPMI since EAMI takes the joint probability into consideration, while EPMI is used to measure the strength degree of a collocation candidate. For example, parse tree re-ranking can be performed by considering the EPMI of the included collocation candidates in parse trees.

3.3. Collocation relationship between any two words

Given any two words w_i and w_j, the EPMI and EAMI between them are defined as the EPMI and EAMI of the optimal collocation candidate related to the two words. Here, the optimal collocation candidate is determined by maximizing the EPMI among all the related collocation candidates over all possible word and feature bigrams and all the possible collocation relation types:

$$EPMI(w_i, w_j) = \max_{\text{all } CR_k,\, wf_i,\, \text{and } wf_j} \{EPMI(wf_i - CR_k - wf_j)\} \tag{11}$$

$$EAMI(w_i, w_j) = EAMI(wf_i - CR_k - wf_j). \tag{12}$$

Here in Equation (12), the collocation candidate $wf_i - CR_k - wf_j$ is determined through maximizing $EPMI(w_i, w_j)$ in Equation (11).

4. Building a Collocation Net

Given a large raw corpus and a general-purpose full parser, a collocation net can be built iteratively as follows:

(1) Parse all the sentences in the large raw corpus into parsed trees using a general-purpose full parser. For every sentence, the N-best (e.g. N = 20) parsed tree hypotheses (PTHs) are kept and their relative probabilities are computed as $\overline{P}_{T_{ij}} = P_{T_{ij}} / \sum_{j=1}^{N_{PTH}} P_{T_{ij}}$.

(2) Extract collocation candidates via linguistic analysis as follows:

(a) First, all the possible collocation candidates are extracted from the PTHs (see Section 5 for details) and their frequencies are accumulated. Assume T_i to be the set of the N-best PTHs for the ith sentence in the corpus and T_{ij} the jth PTH in T_i, the frequency of a collocation candidate in T_{ij} is equal to the relative probability of T_{ij} in T_i, i.e. $\overline{P}_{T_{ij}}$.

(b) Then, $f(wf_i - CR_k - wf_j)$, the frequency of the collocation candidate $wf_i - CR_k - wf_j$, is summed over all the PTHs where $wf_i - CR_k - wf_j$ occurs. In this way, we have a large set of collocation candidates with their frequencies.

(3) Build the collocation net based on the extracted collocation candidates: Given the previously extracted collocation candidates, the collocation net is built by:

(a) First build the first level $L_1 = \{\langle C_{1i}, FDCC_{C_{1i}} \rangle\}$ through the statistics of the extracted collocation candidates: (i) Set $C_1 = \{C_{1i}, 1 \le i \le N_{C_1}\}$, where $C_{1i} = \{wf_i\}$; (ii) Accumulate $FDCC_{C_{1i}} (1 \le i \le |C_1|)$ (the frequency distribution of collocation candidates related to C_{1i}, i.e. wf_i) against various collocation relation types and each class in C_1.

(b) Then construct the second level $L_2 = \{\langle C_{2i}, FDCC_{C_{2i}} \rangle\}$ using the k-means clustering algorithm: (i) Cluster similar classes in the first level into classes in the second level using $FDCC_{C_{1i}} (1 \le i \le |C_1|)$; (ii) Accumulate $FDCC_{C_{2i}} (1 \le i \le |C_2|)$ (the frequency distribution of collocation candidates related to C_{2i}) against various collocation relation types and each class in C_1.

(c) Finally map $FDCC_{C_{1i}} (1 \le i \le |C_1|)$ and $FDCC_{C_{2i}} (1 \le i \le |C_2|)$ against various collocation relation types and each class in C_2 to achieve better granity. Evaluation on full parsing re-ranking in this paper shows that mapping $FDCC_{C_{1i}} (1 \le i \le |C_1|)$ and $FDCC_{C_{2i}} (1 \le i \le |C_2|)$ against each class in C_2 can achieve better performance than C_1.

(4) Examine whether the collocation net is to be re-built. For example, whether the average probability ratio between the best parsed tree hypothesis and the second best parsed tree hypothesis for each sentence converges or begins to drop[d]. If yes, exit the building process. If no, re-build the collocation net by adjusting the probability of each PTH and going to Step (2). For each PTH, e.g. T_{ij} with the original probability $P_{T_{ij}}$, the probability of the PTH can be adjusted by considering the contribution of its included collocation candidates:

$$\log P_{T_{ij}} = (1-a)\log P_{T_{ij}} + a\sum_{i=1}^{|CC|} EPMI(CC_i) \tag{13}$$

where $0 \le a \le 1$ measures the contribution of collocation candidates CC in the PTH (a is fine-tuned to 0.4 using cross-validation) Here, $CC = \{CC_i, 1 \le i \le |CC|\}$ includes all the collocation candidates extracted

[d] In this paper, the threshold for the average probability ratio is set to 0.99.

from T_{ij}; $|CC|$ is the number of collocation candidates extracted from T_{ij} and $EPMI(CC_i)$ is the estimated pair-wise mutual information, which measures the change of information when the collocation candidate CC_i is collocated.

5. Experimentation

The experimentation has been done on the Reuters corpus, which contains 21578 news documents of 2.7 million words in the XML format. In the k-means clustering algorithm, k is fine-tuned to 1000 to achieve proper granity and the frequency distributions of $FDCC_{C_{1i}} (1 \le i \le |C_1|)$ and $FDCC_{C_{2i}} (1 \le i \le |C_2|)$ are mapped to each class in C_2 of the two-level collocation net using cross-validation in this paper. Moreover, the Collins' parser [4] trained on PENN TreeBank is applied and all the collocations are extracted between the head and one modifier of a phrase. In our experimentation, only six most frequently occurring collocation relation types are considered. Table 1 shows them with their occurrence frequencies in the Reuters corpus.

Table 1. Six most frequently occurring collocation relation types (in predicate + argument/adjunct or head noun + modifier format).

Collocation Relation Type	Remark	Freq
VERB-SUB	The right noun is the subject of the left verb	37547
VERB-OBJ	The right noun is the object of the left verb	59124
VERB-PREP	The right preposition modifies the left verb	80493
NOUN-PREP	The right preposition modifies the left noun	19808
NOUN-NOUN	The right noun modifies the left noun	109795
NOUN-ADJ	The right adjective modifies the left noun	139712

To demonstrate the performance of the collocation net, the N-best collocations are extracted from the collocation net. This can be easily done through computing the EAMI and EPMI of all the collocation candidates extracted from the corpus, as described in Section 3. Then all the collocation candidates whose EPMIs are larger than a threshold (e.g. 0) are kept as collocations and sorted according to their EPMIs. As a result, 31,965 collocations are extracted from the Reuters corpus. Table 2 gives some of the examples. It shows that our method can not only extract the collocations that occur frequently in the corpus but also extract the collocations that seldom occur in the corpus. Another advantage is that our method can determine the collocation relationship between any two words and measure its strength degree. In this way, our method can even extract collocations that never occur in the corpus. Table 3 gives some of them. For example, the collocation candidate NOUN(abatement)_NOUN-ADJ_ADJ(eligible) can be measured as a collocation with EAMI of 1.01517e-05 and EPMI of 1.174579 although this collocation candidate does not exist in the corpus. The main reason is that the collocation net provides a word-clustering mechanism to resolve the problem of data sparseness. This is done by using the word-clustering mechanism in the collocation net as shown in Section 3. Table 4 shows an example class "finance/tax" in the second level of the collocation net.

In order to further evaluate the usefulness of the collocation net, we have used it in full parsing re-ranking using the standard PARSEVAL metrics. Here, Collins' parser is used with the standard setting (Sections 2-21 as training data,

Table 2. Examples of N-best collocations (Here, the collocations are sorted according to EPMI first and then EAMI.)

No.	Left Side	Relation Type	Right Side	EPMI	EAMI	Freq
1	NOUN(complex)	NOUN-ADJ	ADJ (aidsrelated)	10.8	0.00023	3
2	NOUN(fraction)	NOUN-ADJ	ADJ(tiny)	10.7	0.00023	3
3	NOUN(politician)	NOUN-ADJ	ADJ(veteran)	10.5	0.00029	3
1001	NOUN(publishing)	NOUN-ADJ	ADJ(desktop)	6.22	0.00045	8
1002	VERB(start)	VERB-SUB	NOUN(talk)	6.22	0.00049	2
1003	NOUN(science)	NOUN-ADJ	ADJ(political)	6.21	.000040	9
5001	VERB(give)	VERB-OBJ	NOUN (breakdown)	3.94	0.00073	11
5002	VERB(introduce)	VERB-OBJ	NOUN(tax)	3.94	0.00018	3
5003	NOUN(fund)	NOUN-NOUN	NOUN(trust)	3.94	0.00051	11
10001	VERB(cut)	VERB-OBJ	NOUN(cost)	2.69	0.00170	11
10002	NOUN(session)	NOUN-NOUN	NOUN(house)	2.69	0.00007	3
10003	NOUN(challenge)	NOUN-PREP	PREP(of)	2.68	0.00147	6
15001	NOUN(factor)	NOUN-ADJ	ADJ(seasonal)	1.85	0.00009	16
15002	NOUN(report)	NOUN-NOUN	NOUN (acreage)	1.85	0.00009	3
15003	NOUN(menu)	NOUN-PREP	PREP(of)	1.85	0.00024	5
20001	NOUN(investor)	NOUN-ADJ	ADJ (Norwegian)	1.20	0.00007	2
20002	NOUN(conflict)	NOUN-ADJ	ADJ(serious)	1.20	0.00001	5
20003	NOUN(country)	NOUN-PREP	PREP(in)	1.20	0.00198	50
31963	NOUN(infusion)	NOUN-PREP	PREP(into)	5^{e}-4	8.27e-9	3
31964	VERB(ask)	VERB-OBJ	NOUN(leader)	4^{e}-4	1.70e-8	3
31965	VERB(land)	VERB-SUB	NOUN(plane)	2e-4	3.19e-8	4

Table 3. Examples of collocations not existing in the corpus

Left Side	Relation Type	Right Side	EPMI	EAMI
NOUN(accountant)	NOUN-ADJ	ADJ(associate)	3.22	8.68e-06
NOUN(worker)	NOUN-NOUN	NOUN(professional)	2.89	1.12e-04
VERB(regain)	VERB-SUB	NOUN(ability)	2.21	8.41e-05
NOUN(stock)	NOUN-ADJ	ADJ(borrowed)	2.12	8.61e-05
NOUN(share)	NOUN-NOUN	NOUN(Malaysia)	1.89	8.65e-05
NOUN(activity)	NOUN-NOUN	NOUN(buying)	1.27	8.71e-05
NOUN(business)	NOUN-NOUN	NOUN(customer)	1.22	9.66e-05
NOUN(abatement)	NOUN-ADJ	ADJ(eligible)	1.17	1.02e-05
VERB(transfer)	VERB-SUB	NOUN(business)	1.06	5.18e-05

Table 4. An example class (“finance/tax”) in the 2nd level of the collocation net.

NOUN(asbestos) NOUN(abatement) NOUN(abba) NOUN(market) NOUN(share) NOUN(stock) NOUN(tax) NOUN(currency) NOUN(contract) NOUN(income) NOUN(reserve) NOUN(investment) NOUN(bid) NOUN(trade) ……

Section 24 as development data and Section 23 as testing data) while 20-best parse trees for each sentence are considered in re-ranking. This is done by building a collocation net on the golden parse trees in the training data and adjusting the probability of each parse tree candidate using the collocation net to achieve full parsing re-ranking, same as Equation (13) applied in Section 4. That is, for each parse tree candidate, e.g. T_{ij} with the original probability $P_{T_{ij}}$, the probability of the parse tree candidate can be adjusted by considering the contribution of its included collocation candidates (Equation (13) is listed here for easy reference):

$$\log P_{T_{ij}} = (1-a)\log P_{T_{ij}} + a\sum_{i=1}^{|CC|} EPMI(CC_i) \qquad (13)$$

where $0 \leq a \leq 1$ measures the contributions of collocation relationship in full parsing re-ranking, $CC = \{CC_i, 1 \leq i \leq |CC|\}$ includes all the collocation candidates extracted from T_{ij}; $|CC|$ is the number of collocation candidates extracted from T_{ij} and $EPMI(CC_i)$ is the estimated pair-wise mutual information, which measures the change of information when the collocation candidate CC_i is collocated. Then, all the parse tree candidates for a sentence can be re-ranked according to their adjusted probabilities as calculated in Equation (13).

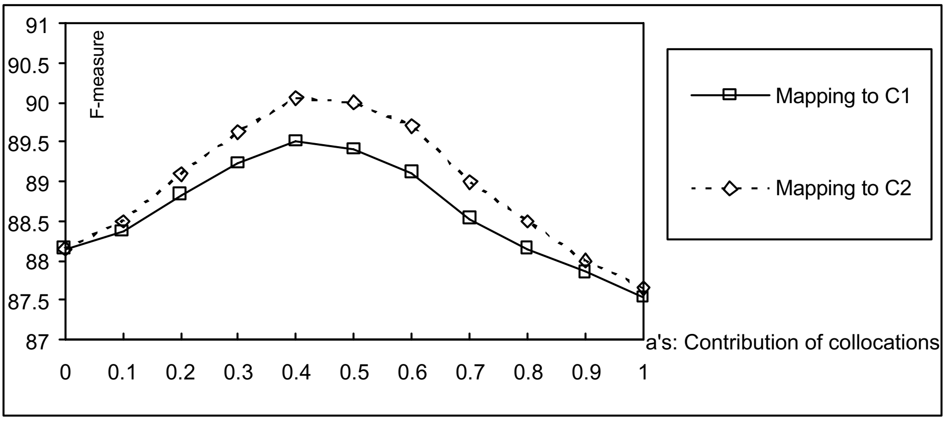

Figure 1. Contributions of collocations in full parsing re-ranking.

Figure 1 compares the effect of different a's in full parsing re-ranking. It shows that the full parsing re-ranking achieves best performance when a = 0.4. Figure 1 also compares the effect of mapping $FDCC_{C_{1i}} (1 \le i \le |C_1|)$ and $FDCC_{C_{2i}} (1 \le i \le |C_2|)$ each class in C_1 and C_2. It shows that mapping them to C_2 achieves better performance than C_1. This suggests the data sparseness problem in building a collocation net and proper handling can improve the performance.

For clarity, Table 5 lists the effect of the best full parsing re-ranking system using the collocation net. It shows that the use of the collocation net can increase the F-measure by 1.9 in F-measure.

Table 5. Application of the collocation net in parse tree re-ranking.

	P(%)	R(%)	F1
Before re-ranking	88.26	88.05	88.15
After re-ranking	90.12	89.98	90.06

6. Conclusion

This paper proposes a novel structure of two-level collocation net and a method capable of automatically building the collocation net given a large raw corpus. Through the collection net, the collocation relationship between any two words can be calculated quantitatively using novel estimated average mutual information (EAMI) as the selection criterion and estimated pair-wise mutual information

(EPMI) as the strength degree. Obviously, the two-level collocation net can be easily extended to more levels through cascading such a two-level structure.

Future works include systematic evaluation of the collocation net on a much larger corpus, its application to other languages such as Chinese and in a general-purpose parser for adaptation to a new domain/application, and development of a more-level hierarchical collocation net.

Acknowledgments

This research is supported by Project 60673041 under the National Natural Science Foundation of China and Project 2006AA01Z147 under the "863" National High-Tech Research and Development of China

References

[1] K. W. Church and H. Patrick, Word association norms, mutual information and lexicography, *ACL'1989*, 1989, pp. 76–83.

[2] K. W. Church and A. G. William, A comparison of the enhanced good turing and deleted estimation methods for estimating probabilities of English bigrams, *Computer, Speech and Language*, 5(1), 1991, 19–54.

[3] K. W. Church and L. M. Robert, Introduction to special issue on computational linguistics using large corpora, *Computational Linguistics*, 19(1), 1993, 1–24.

[4] M. Collins, Head-driven statistical models for natural language parsing, Ph.D. Dissertation, University of Pennsylvania, 1999.

[5] T. Dunning, Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics*, 19(1), 1993, 61–74.

[6] M. Halliday, Lexis as a linguistic level, *in Memory of J. R. Firth,* edited by C. Bazell et al., Longman, 1966.

[7] D. Hindle and M. Rooth, Structural ambiguity and lexical relations, *Computational Linguistics*, 19(1), 1993, 102–119.

[8] J. S. Justeson and S. M. Katz, Technical terminology: Some linguistic properties and an algorithm for identification in text, *Natural Language Engineering*, 1(1), 1995, 9–27.

[9] K. Julian, J. Pederson and F. Chen, A trainable document summarizer, *SIGIR'1995*, 1995, pp. 68–73.

[10] C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999, p. 185.

[11] D. Meyer et al., Loci of contextual effects on visual word recognition, *in Attention and Performance V*, edited by P. Rabbitt and S. Dornie. Academic Press, 1975, pp. 98–116.

[12] I. C. Ross and J. W. Tukey, Introduction to these volumes, in John Wilder Tukey (ed.), *Index to Statistics and Probability*, R&D Press, Los Altos, 1975, pp. iv-x.

[13] R. Rosenfeld, Adaptive statistical language modeling: A maximum entropy approach, Ph.D. Thesis, Carnegie Mellon University, 1994.

[14] F. Smadja, Retrieving collocations from text: Xtract, *Computational Linguistics*, 19(1), 1993, 143–177.

[15] G. W. Snedecor and G. C. William, *Statistical Methods*, Iowa State University Press, Ames, Iowa, 1989, p. 127.

[16] J. Yang, Towards the automatic acquisition of lexical selection rules, *MT Summit VII,* Singapore, 1999, pp. 397–403.

[17] D. Yuret, Discovery of linguistic relations using lexical attraction*,* Ph.D Thesis, cmp-lg/9805009, MIT, 1998.

[18] J. Zhao and C. N. Huang, Aquasi-dependency model for the structural analysis of Chinese BaseNPs, *COLING-ACL'1998,* University de Montreal, Canada, 1998, pp. 1–7.

[19] G. D. Zhou and K. T. Lua, Word association and MI-trigger-based language modeling, *COLING-ACL'1998,* University of Montreal, Canada, 1998, pp. 1465–1471.

[20] G. D. Zhou and K. T. Lua, Interpolation of N-gram and MI-based trigger pair language modeling in mandarin speech recognition, *Computer, Speech and Language*, 13(2), 1999, 123–135.