HYPOTHESIZING WORD ASSOCIATION FROM UNTAGGED TEXT

Tomoyoshi Matsukawa

BBN Systems and Technologies
70 Fawcett St.
Cambridge, MA 02138

ABSTRACT

This paper reports a new method for suggesting word associations, based on a greedy algorithm that employs Chi-square statistics on joint frequencies of pairs of word groups compared against chance co-occurrence. The benefits of this new approach are: 1) we can consider even low frequency words and word pairs, and 2) word groups and word associations can be automatically generated. The method provided 87% accuracy in hypothesizing word associations for unobserved combinations of words in Japanese text.

1. INTRODUCTION

Using mutual information for measuring word association has become popular since [Church and Hanks, 1990] defined word association ratio as mutual information between two words. Word association ratios are a promising tool for lexicography, but there seem to be at least two limitations to the method: 1) much data with low frequency words or word pairs cannot be used and 2) generalization of word usage still depends totally on lexicographers.

In this paper, we propose an alternative (or extended) method for suggesting word associations using Chi-square statistics, which can be viewed as an approximation to mutual information. Rather than considering significance of joint frequencies of word pairs as [Church and Hanks, 1990] did, our algorithm uses joint frequencies of pairs of word **groups** instead. The algorithm employs a hill-climbing search for a pair of word groups that occur significantly frequently.

The benefits of this new approach are:

1) that we can consider even low frequency words and word pairs, and

2) that word groups or word associations can be automatically generated, namely automatic hypothesis of word associations, which can later be reviewed by a lexicographer.

3) word associations can be used in parsing and understanding natural language, as well as in natural language generation [Smadja and McKeown, 1990].

Our method proved to be 87% accurate in hypothesizing word associations for unobserved combinations of words in Japanese text, where accuracy was tested by human verification of a random sample of hypothesized word pairs. We extracted 14,407 observations of word co-occurrences, involving 3,195 nouns and 4,365 verb/argument pairs. Out of this we hypothesized 7,050 word associations. The corpus size was 280,000 words. We would like to apply the same approach to English.

2. RELATED WORK

Some previous work (e.g., [Weischedel, et al., 1990]) found verb-argument associations from bracketed text, such as that in TREEBANK; however, this paper, and related work has hypothesized word associations from untagged text.

[Hindle 1990] confirmed that word association ratios can be used for measuring similarity between nouns. For example, "ship", "plane", "bus", etc., were automatically ranked as similar to "boat". [Resnik 1992] reported a word association ratio for identifying noun classes from a pre-existing hierarchy as selectional constraints on the object of a verb.

[Brown et.al. 1992] proves that, under the assumption of a bi-gram class model, the perplexity of a corpus is minimized when the average mutual information between word classes is maximized. Based on that fact, they cluster words via a greedy search algorithm which finds a local maximum in average mutual information.

Our algorithm considers joint frequencies of pairs of word groups (as [Brown et. al. 1992] does) in contrast to joint frequencies of word pairs as in [Church and Hanks, 1990] and [Hindle 1990]. Here a word group means any subset of the whole set of words. For example, "ship," "plane," "boat" and "car" may be a word group. The algorithm will find pairs of such word groups. Another similarity to [Brown et. al. 1992]'s clustering algorithm is the use of greedy search for a pair of word groups that occur significantly frequently, using an evaluation function based on mutual information between classes.

On the other hand, unlike [Brown et. al. 1992], we assume some automatic syntactic analysis of the corpus, namely part-of-speech analysis and at least finite-state approximations to syntactic dependencies. Moreover, the clustering is done depth first, not breadth first as [Brown et.

al. 1992], i.e., clusters are hypothesized one by one, not in parallel.

3. OVERVIEW OF THE METHOD

The method consists of three phases:

1) **Automatic part of speech tagging of text.** First, texts are labeled by our probabilistic part of speech tagger (POST) which has been extended for Japanese morphological processing [Matsukawa et. al. 1993]. This is fully automatic; human review is not necessary under the assumption that the tagger has previously been trained on appropriate text [Meteer et. al. 1991][1]

2) **Finite state pattern matching.** Second, a finite-state pattern matcher with patterns representing possible grammatical relations, such as verb/argument pairs, nominal compounds, etc. is run over the sample text to suggest word pairs which will be considered candidates for word associations. As a result, we get a word co-occurrence matrix. Again, no human review of the pattern matching is assumed.

3) **Filtering/Generalization of word associations via Chi-square.** Third, given the word co-occurrence matrix, the program starts from an initial pair of word groups (or a submatrix in the matrix), incrementally adding into the submatrix a word which locally gives the highest Chi-square score to the submatrix. Finally, words are removed which give a higher Chi-square score by their removal. By adding and removing words until reaching an appropriate significance level, we get a submatrix as a hypothesis of word associations between the cluster of words represented as rows in the submatrix and the cluster of words represented as columns in the submatrix.

4 WORD SEGMENTATION AND PART OF SPEECH LABELING

Since in Japanese word separators such as spaces are not present, words must be segmented before we assign part of speech to words. To do this, we use JUMAN from Kyoto University to segment Japanese text into words, AMED, an example-based segmentation corrector, and a Hidden Markov Model (POST) [Matsukawa, et. al. 1993]. For example, POST processes an input text such as the following:

一方、国際協会は住友銀行系の独占状態を崩し、資金量の豊富な他の上位都市銀行の系列カード会社と提携することで、日本でのマスターカードとの取扱高逆転を狙っている。

and produces tagged text such as: [2]

一方/CONJ 、/TT 国際/CN 協会/CN は/TM
住友銀行系/PN の/NCM 独占/SN 状態/CN を/CM
崩し/ADV 、/TT 資金/CN 量/CN の/NCM 豊富な/ADJ
他/CN の/NCM 上位/CN 都市/CN 銀行/CN の/NCM
系列/CN カード/CN 会社/CN と/PT 提携/SN する/VB
こと/FN で/PT 、/TT 日本/PN で/PT の/NCM
マスター/CN カード/CN と/PT の/NCM 取扱高逆転/CN
を/CM 狙って/VB いる/VSUF 。/KT

5. FINITE STATE PATTERN MATCHING

We use the following finite state patterns for extracting possible Japanese verb/argument word co-occurrences from automatically segmented and tagged Japanese text. Completely different patterns would be used for English.

$$\begin{Bmatrix} \text{CN} \\ \text{PN} \\ \text{SN} \end{Bmatrix} \begin{Bmatrix} \text{CM} \\ \text{PT} \end{Bmatrix} \cdots \begin{Bmatrix} \text{VB} \\ \text{SN} \end{Bmatrix}$$

where CN = common noun
PN = proper name
SN = Sa-inflection noun (nominal verb)
CM = case marker (-nom/-acc argument)
PT = particle (other arguments)
VB = verb

Here, the first part (CN, PN or SN) represents a noun. Since in Japanese the head noun of a noun phrase is always at the right end of the phrase, this part should always match a head noun. The second part (CM or PT) represents a postposition which identifies an argument of a verb. The final pattern element (VB or SN) represents a verb. Sa-inflection nouns (SN) are nominalized verbs which form a verb phrase with the morpheme "suru."

[1] In our experience thus far in three domains and in both Japanese and English, while retraining POST on domain-specific data would reduce the error rate, the effect on overall performance of the system in data extraction from text has been small enough to make retraining unnecessary. The effect of domain-specific lexical entries (e.g., DRAM is a noun in microelectronics) often mitigates the need to retrain.

[2] CONJ = conjunction; TT = Japanese comma; CN = common noun; TM = Topic marker; PN - proper noun; etc.

Distance	Matched Text
0	系列/CN カード/CN 会社/CN と/PT 提携/SN する/VBこと/FN で/PT 、...
1	フランス/PN の/NCM 会社/CN と/PT 技術/CN 提携/SNを/CM して/VB ...
2	証券/ADJ 会社/CN と/PT デパート/CN が/CM 提携/SNして/VB ...
4	大手/CN 証券/CN 会社/CN と/PT 銀行/CN 系/NNSUカード/CN の/NCM 提携/SN の/NCM 動き/CN

Figure 1: *Examples of Pattern Matches with Skipping over Words.*

Since argument structure in Japanese is marked by postpositions, i.e., case markers (i.e., "o," "ga") and particles (e.g., "ni," "kara," . . .), word combinations matched with the patterns will represent associations between a noun filling a particular argument type (e.g., "o") and a verb. Note that topic markers (TM; i.e., "wa") and toritate markers (TTM; e.g."mo", "sae", ...) are not included in the pattern since these do not uniquely identify the case of the argument.

Just as in English, the arguments of a verb in Japanese may be quite distant from the verb; adverbial phrases and scrambling are two cases that may separate a verb from its argument(s). We approximate this in a finite state machine by allowing words to be skipped. In our experiment, up to four words could be skipped. As shown in Figure 1, matching an argument structure varies from distance 0 to 4.

By limiting the algorithm to a maximum of four word gaps, and by not considering the ambiguous cases of topic markers and taritate markers, we have chosen to limit the cases considered in favor of high accuracy in automatically hypothesizing word associations. [Brent, 1991] similarly limited what his algorithm could learn in favor of high accuracy.

6. FILTERING AND GENERALIZATION VIA CHI-SQUARE

Word combinations found via the finite state patterns include a noun, postposition, and a verb. A two dimensional matrix (a word co-occurrence matrix) is formed, where the columns are nouns, and the rows are pairs of a verb plus postposition. The cells of the matrix are the frequency of the noun (column element) co-occurring in the given case with that verb (row element).

Starting from a submatrix, the algorithm successively adds to the submatrix the word with the largest Chi-square score among all words outside the submatrix. Words are added until a local maximum is reached. Finally, the appropriateness of the submatrix as a hypothesis of word associations is checked with heuristic criteria based on the sizes of the row and the column of the submatrix. Currently, we use the following criteria for appropriateness of a submatrix:

LET l : size of row of submatrix
m : size of column of submatrix
C1, C2, C3 : parameters
IF l > C1, and
m > C1, and
l > C2 or m/l < C3, and
m > C2 or l/m < C3
THEN the submatrix is appropriate.

For any submatrix found, the co-occurrence observations for the clustered words are removed from the word co-occurrence matrix and treated as a single column of clustered nouns and a single row of clustered verb plus case pairs. Currently, we use the following values for the parameters: C1=2, C2=10, and C3=10.

Table 1. shows an example of clustering starting from the initial submatrix shown in Figure 2. The words in Figure 2 were manually selected as words meaning "organization." In Table 1, the first (leftmost) column indicates the word which was added to the submatrix at each step. The second column gives an English gloss of the word. The third column reports f(x,Y), the frequency of the co-occurrences between the word and the words that co-occur with it. For example, the first line of the table shows that the word "を／設立" (establish/-acc) co-occurred with the "organization" words 26 times. The rightmost column specifies I(X,Y), the scaled mutual information between the rows and columns of the submatrix. As the clustering proceeds, I(X,Y) gets larger.

会社(company), 本社(head quarter), 機関(organization), 企業(coorporation), 両社(both companies), 学校(school), 同社(the company), 子会社(child company), 銀行(bank), 百貨店(department store), 代理店(agency), 生協(coop.), 商社(business company), 都銀(city bank), 売店(stand), 信託銀行(trust bank), 支店(branch), 信金(credit association), 本店(head store), 大学(university), 各社(each company), デパート(department store), 農協(agriculture cooperative), メーカー(maker), 書店(book store), テレビ局(TV station), プロダクション(agency), スーパー(supermarket), 株式会社(joint-stock corporation), 医院(doctor's office), 全店(all stores)

Figure 2: *The initial word group (submatrix) for the clustering shown in Table 1.*

Word added	Gloss	Freq	I
を／設立	establish/-acc	26	0.11
と／提携	tie-up/with	25	0.19
が／提携	tie-up/-nom	18	0.25
と／組む	unite/with	11	0.29
が／協力	cooperate/-nom	7	0.32
が／持つ	possess/-nom	8	0.35
が／組む	unite/-nom	7	0.38
が／進出	advance/-nom	6	0.40
と／相次ぐ	in succession	5	0.43
が／進める	proceed/-nom	4	0.44
を／買収	purchase/-acc	5	0.46
に／委託	entrust/-acc	6	0.47
が／生産	produce/-nom	6	0.49
が／開発	develop/-nom	7	0.51
が／出資	invest/-nom	6	0.52
と／拡大	expand/with	3	0.54
が／共同開発	develop/-nom	3	0.55
が／発行	publish/-nom	4	0.56
が／合意	agree/-nom	3	0.58
に／求める	demand/from	3	0.59
に／出資	invest/in	5	0.60
が／販売	sell/-nom	7	0.61
が／買収	purchase/-nom	3	0.63
を／開設	open/-acc	4	0.64
から／導入	introduce/from	3	0.65
が／つくる	create/-nom	3	0.66
で／使う	utilize/at	3	0.67
に／限る	limit/to	3	0.68
が／扱う	treat/-nom	3	0.69
が／結ぶ	connect/-nom	3	0.69
が／する	do/-nom	5	0.70
を／除く	exclude/-acc	3	0.71
に／対抗	oppose/to	3	0.71
で／調印	sign/-copula	3	0.72
に／販売	sell/to	4	0.72
に／参加	participate/in	4	0.72
法人	corporation	9	0.74
大手	major	5	0.75
ジャパン	Japan	5	0.77
日商岩井	Nisho-Iwai	4	0.78
三者	three parties	3	0.79
製薬	Drug Company	4	0.80
ソニー	Sony	5	0.81
業者	dealer	5	0.81
研究所	Institution	5	0.82
本田	Honda	4	0.83
三菱重工	Mitsubishi	3	0.83
ＡＴＴ	AT&T	3	0.84
航	Air Line	4	0.84
それぞれ	respectively	3	0.85
本田技研	Honda	3	0.85
銀	Bank	7	0.85
航空	Air Line	6	0.85
信託	Trust Company	4	0.85
鉄	Steel Company	4	0.85

Table 1: *Example of Clustering*

7. EVALUATION

Using 280,000 words of Japanese source text from the TIPSTER joint ventures domain, we tried several variations of the initial submatrices (word groups) from which the search in step three of the method starts:

a) complete bipartite subgraphs,
b) pre-classified noun groups and
c) significantly frequent word pairs.

Based on the results of the experiments, we concluded that alternative (b) gives both the most accurate word associations and the highest coverage of word associations. This technique is practical because classification of nouns is generally much simpler than that of verbs. We don't propose any automatic algorithm to accomplish noun classification, but instead note that we were able to manually classify nouns in less than ten categories at about 500 words/hour. That productivity was achieved using our new tool for manual word classification, which is partially inspired by EDR's way of classifying their semantic lexical data [Matsukawa and Yokota, 1991].

Based on a corpus of 280,000 words in the TIPSTER joint ventures domain, the most frequently occurring Japanese nouns, proper nouns, and verbs were automatically identified. Then, a student classified the frequently occurring nouns into one of the twelve categories in (1) below, and each frequently occurring proper noun into one of the four categories in (2) below, using a menu-based tool, we were able to categorize 3,195 lexical entries in 12 person-hours.[3] These categories were then used as input to the word co-occurrence algorithm.

1. Common noun categories
1a. Organization
CORPORATION
GOVERNMENT
UNDETERMINED-CORPORATION
OTHER-ORGANIZATION
1b. Location
CITY
COUNTRY
PROVINCE

[3] We divided the process of classifying common nouns into two phases; classification into the four categories 1a, 1b, 1c and 1d, and further classification into the twelve categories. As a result, each word was checked twice. We found that using two phases generally improves both overall productivity and consistency.

OTHER-LOCATION
1c. Person
ENTITY-OFFICER
TITLE
OTHER-PERSON
1d. Other
2. Proper noun categories
ORGANIZATION
LOCATION
PERSON
OTHER

Using the 280,000 word joint venture corpus, we collected 14,407 word co-occurrences, involving 3,195 nouns and 4,365 verb/argument pairs, by the finite state pattern given in Section 5. 16 submatrices were clustered, grouping 810 observed word co-occurrences and 6,240 unobserved (or hypothesized) word co-occurrences. We evaluated the accuracy of the system by manual review of a random sample of 500 hypothesized word co-occurrences. Of these, 435, or 87% were judged reasonable. This ratio is fine compared with a random sample of 500 arbitrary word co-occurrences between the 3,195 nouns and the 4,365 verb/argument pairs, of which only 153 (44%) were judged reasonable. Table 2 below shows some examples judged reasonable; questionable examples are marked by "?"; unreasonable hypotheses are marked with an asterisk.

With a small corpus (280,000 words) such as ours, considering small frequency co-occurrences is critical. Looking at Table 3 below, if we had to ignore co-occurrences with frequency less than five (as [Church and Hanks 1990] did), there would be very little data. With our method, as long as the frequency of co-occurrence of the word being considered with the set is greater than two, the statistic is stable.

Frequency	Number of Word Pairs
0	6240
1	631
2	113
3	36
4	18
5	4
6	2
7	3
9	1
10	1
16	1

Table 3: *Pair Frequencies*

8. CONCLUSION

Our method achieved fully automatic hypothesis of word associations, starting from untagged text and generalizing to unobserved word associations. As a result of human review 87% of the hypotheses were judged to be reasonable. Because the technique considers low frequency cases, most of the data was used in making generalizations.

It remains to be determined how well this method will work for English, but with appropriate finite state patterns, similar results may be achieved.

	オーナー	に／就任
	(owner)	(take office/as)
	ＡＴＴ	から／導入
	(AT&T)	(introduce/from)
	首都圏	に／建設
	(metropolitan)	(build/at)
	要員	を／派遣
	(personnel)	(dispatch/-acc)
	委員会	と／組む
	(Commitee)	(unite/with)
	図書館	が／販売
	(library)	(sell/-nom)
	商会	を／結成
	(Company)	(organize/-acc)
	代理店	が／発行
	(agency)	(publish/-nom)
	郵便局	と／提携
	(post office)	(tie-up/with)
	州	に／展開
	(State)	(develop/to)
	キャノン	に／入る
	(Cannon)	(enter/-acc)
	医院	に／限る
	(doctor's office)	(limit/to)
	諸国	に／持つ
	(nations)	(have/in)
?	野村	が／生産
	(Nomura)	(produce/-nom)
?	駅員	が／就任
	(station employee)	(take office/-nom)
*	ＤＲＡＭ	が／結ぶ
	(DRAM)	(unite/-nom)
*	スイス	が／みる
	(Switzerland)	(see/-nom)
*	取締役	に／発表
	(director)	(announce/to)

Table 2: *Examples of reasonable hypothesized co-occurrences*

ACKNOWLEDGMENTS

The author wishes to thank Madeleine Bates, Ralph Weischedel and Sean Boisen for significant contributions to this paper.

REFERENCES

1. Brent, M.R., (1991) "Automatic Acquisition of Subcategorization Frames from Untagged Text," *Proceedings of the 29th annual Meeting of the ACL*, pp. 209-214.
2. Brown, P.F., et. al., (1992) "Class-based N-gram Models of Natural Language," *Computational Linguistics* Vol. 18 (4), pp. 467-479.
3. Church, K. and Hanks, P., (1990) "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics* Vol. 16 (1), pp.22-29.
4. Hindle, D., (1990) "Noun Classification from Predicate-Argument Structures," *Proceedings of the 28th Annual Meeting of the ACL*, pp. 268-275.
5. Hoel P. G., (1971): *Introduction to Mathematical Statistics*, Chapter 9. 2.
6. Resnik, P., (1992) "A Class-based Approach to Lexical Discovery," *Proceedings of the 30th Annual Meeting of the ACL*, pp. 327-329.
7. Smadja F.A. and McKeown, K.R., (1990) "Automatically Extracting and Representing Collocations for Language Generation," *Proceedings of the 28th Annual Meeting of the ACL*, pp. 252-259.
8. Matsukawa T., Miller S. and Weischedel R. (1993) "Example-based Correction of Word Segmentation and Part of Speech Labelling," *Proceedings of DARPA Human Language Technologies Workshop.*
9. Matsukawa , T. and Yokota, E. (1991) "Development of the Concept Dictionary - Implementation of Lexical Knowledge," *Proc. of pre-conference workshop sponsored by the special Interest Group on the Lexicon (SIGLEX) of the Association for Computational Linguistics,* 1991.
10. Weischedel, R. et al. (1991) "Partial Parsing: A Report on Work in Progress," *Proceedings of the Workshop on Speech and Natural Language*, pp. 204-210.

APPENDIX: JUSTIFICATION OF CHI SQUARE

Chi-square score is given by the following formula :

$$I(X, Y) = \sum I(X, Y) = \sum p(X, Y) \log \frac{p(X, Y)}{p(X)\, p(Y)} \quad (0)$$

where X, Y = columns and rows of a word co-occurrence matrix

X, Y = subsets of X, Y, respectively

(i.e. word classes at the columns and the rows)

This can be justified as follows.

According to [Hoel 1971], the likelihood ratio *LAMBDA* for a test of the hypothesis: p(i) = po(i) (i = 1, 2, . . ., k), where p(i) is the probability of case i and po(i) is a hypothesized probability of it, when observations are independent of each other, is given as:

$$-2 \log LAMBDA = 2 \sum_{i=1}^{k} n(i) \log \frac{n(i)}{e(i)} \quad (1)$$

where n(i) is the number of observations of case i, and e(i) is its expectation, i.e., e(i) = n p(i), where n is the total number of observations.

The distribution is chi-square when n is large. If we assume two word classes, ci and cj, occur independently, then the expected value of the probability of their co-occurrence will be,

$$e(c_i, c_j) = n\, p(c_i)\, p(c_j) \quad (2)$$

where p(ci) and p(cj) are estimations of the probability of occurrence of ci and cj. The maximum likelihood estimate of p(ci) and p(cj) is f(ci)/n and f(cj)/n, where f(cj) and f(cj) are the number of observations of words classified in ci and cj. The maximum likelihood estimate of p(ci, cj), the probability of the co-occurrences of words in ci and cj, is f(ci, cj)/n, where f(ci, cj) is the number of observations of the co-occurrences. Then the number of the co-occurrences n(ci, cj) (which is the same as f(ci, cj)) can be represented as,

$$n(c_i, c_j) = n\, p(c_i, c_j) \quad (3)$$

Therefore, given k classes, c1, c2, ..., ck, substituting (2) and (3) into (1).

$$2 \sum_{i=0}^{k} \sum_{j=0}^{i} n\, p(c_i, c_j) \log \frac{p(c_i, c_j)}{p(c_i)\, p(c_j)} \quad (4)$$

If n is large, this will have a chi-square distribution; therefore, we can estimate how unlikely our assumption of independence among word classes is. Since formula (4) gives a scaled average mutual information among the word classes, searching for a partition of words that provides maximum average mutual information among word classes is equivalent to seeking classes where independence among word classes is minimally likely. The algorithm reported in this paper searches for pairs of word classes which provide a local maximum I(X, Y), a term in the summation of formula (0).