Word Association and MI-Trigger-based Language Modeling

GuoDong ZHOU KimTeng LUA
Department of Information Systems and Computer Science
National University of Singapore
Singapore 119260
{zhougd, luakt}@iscs.nus.edu.sg

Abstract

There exists strong word association in natural language. Based on mutual information, this paper proposes a new MI-Trigger-based modeling approach to capture the preferred relationships between words over a short or long distance. Both the distance-independent(DI) and distance-dependent(DD) MI-Trigger-based models are constructed within a window. It is found that proper MI-Trigger modeling is superior to word bigram model and the DD MI-Trigger models have better performance than the DI MI-Trigger models for the same window size. It is also found that the number of the trigger pairs in an MI-Trigger model can be kept to a reasonable size without losing too much of its modeling power. Finally, it is concluded that the preferred relationships between words are useful to language disambiguation and can be modeled efficiently by the MI-Trigger-based modeling approach.

Introduction

In natural language there always exist many preferred relationships between words. Lexicographers always use the concepts of collocation, co-occurrence and lexis to describe them. Psychologists also have a similar concept: word association. Two highly associated word pairs are "not only/but also" and "doctor/nurse". Psychological experiments in [Meyer+75] indicated that the human's reaction to a highly associated word pair was stronger and faster than that to a poorly associated word pair.

The strength of word association can be measured by mutual information. By computing mutual information of a word pair, we can get many useful preference information from the corpus, such as the semantic preference between noun and noun(e.g."doctor/nurse"), the particular preference between adjective and noun(e.g."strong/currency"), and solid structure (e.g."pay/attention")[Calzolori90]. These information are useful for automatic sentence disambiguation. Similar research includes [Church90], [Church+90], Magerman+90], [Brent93], [Hiddle+93], [Kobayashi+94] and [Rosenfeld94].

In Chinese, a word is made up of one or more characters. Hence, there also exists preferred relationships between Chinese characters. [Sproat+90] employed a statistical method to group neighboring Chinese characters in a sentence into two-character words by making use of a measure of character association based on mutual information. Here, we will focus instead on the preferred relationships between words.

The preference relationships between words can expand from a short to long distance. While N-gram models are simple in language modeling and have been successfully used in many tasks, they have obvious deficiencies. For instance, N-gram models can only capture the short-distance dependency within an N-word window where currently the largest practical N for natural language is three and many kinds of dependencies in natural language occur beyond a three-word window. While we can use conventional N-gram models to capture the short-distance dependency, the long-distance dependency should also be exploited properly.

The purpose of this paper is to study the preferred relationships between words over a short or long distance and propose a new modeling approach to capture such phenomena in the Chinese language.

This paper is organized as follows: Section 1 defines the concept of trigger pair. The criteria of selecting a trigger pair are described in Section 2 while Section 3 describes how to measure the strength of a trigger pair. Section 4 describes trigger-based language modeling. Section 5 gives one of its applications: PINYIN-to-Character Conversion. Finally, a conclusion is given.

1 Concept of Trigger Pair

Based on the above description, we decide to use the trigger pair[Rosenfeld94] as the basic concept for extracting the word association information of an associated word pair. If a word A is highly associated with another word B, then $(A \rightarrow B)$ is considered a "trigger pair", with A being the trigger and B the triggered word. When A occurs in the document, it triggers B, causing its probability estimate to change. A and B can be also extended to word sequences. For simplicity, here we will concentrate on the trigger relationships between single words although the ideas can be extended to longer word sequences.

How to build a trigger-based language model? There remain two problems to be solved: 1) how to select a trigger pair? 2) how to measure a trigger pair?

We will discuss them separately in the next two sections.

2 Selecting Trigger Pair

Even if we can restrict our attention to the trigger pair (A, B) where A and B are both single words, the number of such pairs is too large. Therefore, selecting a reasonable number of the most powerful trigger pairs is important to a trigger-based language model.

2.1 Window Size

The most obvious way to control the number of the trigger pairs is to restrict the window size, which is the maximum distance between the trigger pair. In order to decide on a reasonable window size, we must know how much the distance between the two words in the trigger pair affects the word probabilities.

Therefore, we construct the long-distance Word Bigram(WB) models for distance-$d = 1,2,...,100$. The distance-100 is used as a control, since we expect no significant information after that distance. We compute the conditional perplexity[Shannon51] for each long-distance WB model.

Conditional perplexity is a measure of the average number of possible choices there are for a conditional distribution. The conditional perplexity of a conditional distribution with conditional entropy $H(Y|X)$ is defined to be $2^{H(Y|X)}$. Conditional Entropy is the entropy of a conditional distribution. Given two random variables X and Y, a conditional probability mass function $P_{Y|X}(y|x)$, and a marginal probability mass function $P_Y(y)$, the conditional entropy of Y given X, $H(Y|X)$ is defined as:

$$H(Y|X) = -\sum_{x\in X}\sum_{y\in Y} P_{X,Y}(x,y)\log_2 P_{Y|X}(y|x) \qquad (1)$$

For a large enough corpus, the conditional perplexity is usually an indication of the amount of information conveyed by the model: the lower the conditional perplexity, the more information it conveys and thus a better model. This is because the model captures as much as it can of that information, and whatever uncertainty remains shows up in the conditional perplexity. Here, the training corpus is the XinHua corpus, which has about 57M(million) characters or 29M words.

From Table 1 we find that the conditional perplexity is lowest for d = 1, and it increases significantly as we move through d = 2, 3, 4, 5 and 6. For d = 7, 8, 9, 10, 11, the conditional perplexity increases slightly. We conclude that significant information exists only in the last 6 words of the history. However, in this paper we restrict maximum window size to 10.

Distance	Perplexity	Distance	Perplexity
1	230	7	1479
2	575	8	1531
3	966	9	1580
4	1157	10	1599
5	1307	11	1611
6	1410	100	1674

Table 1: Conditional perplexities of the long-distance WB models for different distances

2.2 Selecting Trigger Pair

Given a window, we define two events:

w : { w is the next word }

w_o : { w_o occurs somewhere in the window}

Considering a particular trigger $(A \rightarrow B)$, we are interested in the correlation between the two events A_o and B.

A simple way to assess the significance of the correlation between the two events A_o and B in the trigger $(A \rightarrow B)$ is to measure their cross product ratio(CPR). One often used measure is the logarithmic measure of that quality, which has units of bits and is defined as:

$$\log CPR(A_o, B) = \log \frac{P(A_o, B)P(\overline{A_o}, \overline{B})}{P(A_o, \overline{B})P(\overline{A_o}, B)} \qquad (2)$$

where $P(X_o, Y)$ is the probability of a word pair (X_o, Y) occurring in the window.

Although the cross product ratio measure is simple, it is not enough in determining the utility of a proposed trigger pair. Consider a highly correlated pair consisting of two rare words (树梢 → 白皑皑), and compare it to a less well correlated, but more common pair (医生 → 护士). An occurrence of the word "树梢"(tail of tree) provides more information about the word "白皑皑"(pure white) than an occurrence of the word "医生"(doctor) about the word "护士"(nurse). Nevertheless, since the word "医生" is likely to be much more common in the test data, its average utility may be much higher. If we can afford to incorporate only one of the two pairs into our trigger-based model, the trigger pair(医生 → 护士) may be preferable.

Therefore, an alternative measure of the expected benefit provided by A_o in predicting B is the average mutual information(AMI) between the two:

$$\begin{aligned} AMI(A_o; B) = {} & P(A_o, B)\log \frac{P(A_o B)}{P(A_o)P(B)} \\ & + P(A_o, \overline{B})\log \frac{P(A_o \overline{B})}{P(A_o)P(\overline{B})} \\ & + P(\overline{A_o}, B)\log \frac{P(\overline{A_o} B)}{P(\overline{A_o})P(B)} \\ & + P(\overline{A_o}, \overline{B})\log \frac{P(\overline{A_o}\overline{B})}{P(\overline{A_o})P(\overline{B})} \qquad (3) \end{aligned}$$

Obviously, Equation 3 takes the joint probability into consideration. We use this equation to select the trigger pairs. In related works, [Rosenfeld94] used this equation and [Church+90] used a variant of the first term to automatically identify the associated word pairs.

3 Measuring Trigger Pair

Considering a trigger pair $(A_o \rightarrow B)$ selected by average mutual information $AMI(A_o; B)$ as shown in Equation 3, mutual information $MI(A_o; B)$ reflects the degree of preference relationship between the two words in the trigger pair, which can be computed as follows:

$$MI(A_o; B) = \log \frac{P(A_o, B)}{P(A_o) \cdot P(B)} \qquad (4)$$

where $P(X)$ is the probability of the word X occurred in the corpus and $P(A, B)$ is the probability of the word pair (A, B) occurred in the window.

Several properties of mutual information are apparent:

- $MI(A_o; B)$ is deferent from $MI(B_o; A)$, i.e. mutual information is ordering dependent.
- If A_o and B are independent, then $MI(A; B) = 0$.

In the above equations, the mutual information $MI(A_o; B)$ reflects the change of the *information content* when the two words A_o and B are correlated. This is to say, the higher the value of $MI(A_o; B)$, the stronger affinity the words A_o and B have. Therefore, we use mutual information to measure the preference relationship degree of a trigger pair.

5 MI-Trigger-based Modeling

As discussed above, we can restrict the number of the trigger pairs using a reasonable window size, select the trigger pairs using average mutual information and then measure the trigger pairs using mutual information. In this section, we will describe in greater detail about how to build a trigger-based model. As the triggers are mainly determined by mutual information, we call them MI-Triggers. To build a concrete MI-Trigger model, two factors have to be considered.

Obviously one is the window size. As we have restricted the maximum window size to 10, we will experiment on 10 different window sizes($ws = 1,2,\ldots,10$).

Another one is whether to measure an MI-Trigger in a distance-independent(DI) or distance-dependent(DD) way. While a DI MI-Trigger model is simple, a DD MI-Trigger model has the potential of modeling the word association better and is expected to have better performance because many of the trigger pairs are distance-dependent. We have studied this issue using the XinHua corpus of 29M words by creating an index file that contains. For every word, a record of all of its occurrences with distance-dependent co-occurrence statistics. Some examples are shown in Table 2, which shows that "越／越"("the more/the more") has the highest correlation when the distance is 2, that "不但／而且"("not only/but also") has the highest correlation when the distances are 3, 4 and 5, and that "医生／护士 "("doctor/nurse") has the highest correlation when the distances are 1 and 2. After manually browsing hundreds of the trigger pairs, we draw following conclusions:

- Different trigger pairs display different behaviors.
- Behaviors of trigger pairs are distance-dependent and should be measured in a distance-dependent way.
- Most of the potential of triggers is concentrated on high-frequency words. (医生→护士) is indeed more useful than (树梢→白皑皑).

Distance	越/越	不但/而且	医生/护士
1	0	0	24
2	3848	5	15
3	72	24	1
4	65	18	1
5	45	14	0
6	45	4	0
7	40	2	0
8	23	3	0
9	9	2	1
10	8	4	0

Table 2: The occurrence frequency of word pairs as a function of distance

To compare the effects of the above two factors, 20 MI-trigger models(in which DI and DD MI-Trigger models with a window size of 1 are same) are built. Each model differs in different window sizes, and whether the evaluation is done in the DI or DD way. Moreover, for ease of comparison, each MI-Trigger model includes the same number of the best trigger pairs. In our experiments, only the best 1M trigger pairs are included. Experiments to determine the effects of different numbers of the trigger pairs in a trigger-based model will be conducted in Section 5.

For simplicity, we represent a trigger pair as XX- ws -MI-Trigger, and call a trigger-based model as the XX- ws -MI-Trigger model, while XX represents DI or DD and ws represents the window size. For example, the DD-6-MI-Trigger model represents a distance-dependent MI-Trigger-based model with a window size of 6.

All the models are built on the XinHua corpus of 29M words. Let' s take the DD-6-MI-Trigger model as a example. We filter about $28 \times 28 \times 6$M(with six different distances and with about 28000 Chinese words in the lexicon) possible DD word pairs. As a first step, only word pairs that co-occur at least 3 times are kept. This results in 5.7M word pairs. Then selected by average mutual information, the best 1M word pairs are kept as trigger pairs. Finally, the best 1M MI-Trigger pairs are measured by mutual information. In this way, we build a DD-6-MI-Trigger model which includes the best 1M trigger pairs.

Since the MI-Trigger-based models measure the trigger pairs using mutual information which only reflects the change of information content when the two words in the trigger pair are correlated, a word unigram model is combined with them. Given $S = w_1 w_2 \ldots w_n$, we can estimate the logarithmic probability $\log P(S)$. For a DI- ws MI-Trigger-based model,

$$\log P(S) = \sum_{i=1}^{n} \log P(w_i) + \sum_{i=n}^{2} \sum_{j=i-1}^{\max(1,i-ws)} DI - ws - MI - Trigger(w_j \rightarrow w_i) \qquad (5)$$

and for a DD- ws -MI-Trigger-based model,

$$\log P(S) = \sum_{i=1}^{n} \log P(w_i)$$

$$+\sum_{i=n}^{2}\sum_{j=i-1}^{\max(1,i-ws)} DD-ws-MI-Trigger(w_j \rightarrow w_i, i-j+1) \qquad (6)$$

where ws is the windows size and $i-j+1$ is the distance between the words w_i and w_j. The first item in each of Equation 5 and 6 is the logarithmic probability of S using a word unigram model and the second one is the value contributed to the MI-Trigger pairs in the MI-Trigger model.

In order to measure the efficiency of the MI-Trigger-based models, the conditional perplexities of the 20 different models (each has 1M trigger pairs) are computed from the XinHua corpus of 29M words and are shown in Table 3.

Window Size	Distance - Independent	Distance - Dependent
1	301	301
2	288	259
3	280	238
4	272	221
5	267	210
6	262	201
7	270	216
8	275	227
9	282	241
10	287	252

Table 3: The conditional perplexities of the 20 different MI-Trigger models

5 PINYIN-to-Character Conversion

As an application of the MI-Trigger-based modeling, a PINYIN-to-Character Conversion (PYCC) system is constructed. In fact, PYCC has been one of the basic problems in Chinese processing and the subjects of many researchers in the last decade. Current approaches include:

- The longest word preference algorithm [Chen+87] with some usage learning methods [Sakai+93]. This approach is easy to implement, but the hitting accuracy is limited to 92% even with large word dictionaries.
- The rule-based approach [Hsieh+89] [Hsu94]. This approach is able to solve the related lexical ambiguity problem efficiently and the hitting accuracy can be enhanced to 96%.
- The statistical approach [Sproat92] [Chen93]. This approach uses a large corpus to compute the N-gram and then uses some statistical or mathematical models, e.g. HMM, to find the optimal path through the lattice of possible character transliterations. The hitting accuracy can be around 96%.
- The hybrid approach using both the rules and statistical data[Kuo96]. The hitting accuracy can be close to 98%.

In this section, we will apply the MI-Trigger-based models in the PYCC application. For ease of comparison, the PINYIN counterparts of 600 Chinese sentences(6104 Chinese characters) from Chinese school text books are used for testing.

The PYCC recognition rates of different MI-Trigger models are shown in Table 4.

Window Size	Distance - Independent	Distance - Dependent
1	93.6%	93.6%
2	94.4%	95.5%
3	94.7%	96.1%
4	95.0%	96.3%
5	95.2%	96.5%
6	95.3%	96.6%
7	94.9%	96.4%
8	94.6%	96.2%
9	94.5%	96.1%
10	94.3%	95.8%

Table 4: The PYCC recognition rates for the 20 MI-Trigger models

No. of the MI-Trigger Pairs	Perplexity	Recognition Rate
0	1967	85.3%
100,000	672	90.7%
200,000	358	92.6%
400,000	293	94.2%
600,000	260	95.5%
800,000	224	96.3%
1,000,000	201	96.6%
1,500,000	193	96.9%
2,000,000	186	97.2%
3,000,000	183	97.2%
4,000,000	181	97.3%
5,000,000	178	97.6%
6,000,000	175	97.7%

Table 5: The effect of different numbers of the trigger pairs on the PYCC recognition rates

Table 4 shows that the DD-MI-Trigger models have better performances than the DI-MI-Trigger models for the same window size. Therefore, the preferred relationships between words should be modeled in a DD way. It is also found that the PYCC recongition rate can reach up to 96.6%.

As it was stated above, all the MI-Trigger models only include the best 1M trigger pairs. One may ask: what is a reasonable number of the trigger pairs that an MI-Trigger model should include? Here, we will examine the effect of different numbers of the trigger pairs in an MI-Trigger model on the PINYIN-to-Character conversion rates. We use the DD-6-MI-Trigger model and the result is shown in Table 5.

We can see from Table 5 that the recognition rate rises quickly from 90.7% to 96.3% as the number of MI-Trigger pairs increases from 100,000 to 800,000 and then it rises slowly from 96.6% to 97.7% as the number of MI-Triggers increases from 1,000,000 to 6,000,000. Therefore, the best 800,000 trigger pairs should at least be included in the DD-6-MI-Trigger model.

Model	Word Unigram	Word Bigram	DD-6-MI-Trigger
Parameter Numbers	28,000	$28{,}000^2$ $\approx 7.8 \times 10^8$	$5 \times 10^6 + 28{,}000$ $\approx 5.0 \times 10^6$
Perplexity	1967	230	178

Table 6: Comparison of word unigram, bigram and MI-Trigger model

In order to evaluate the efficiency of MI-Trigger-based language modeling, we compare it with word unigram and bigram models. Both word unigram and word bigram models are trained on the XinHua corpus of 29M words. The result is shown in Table 6. Here the DD-6-MI-Trigger model with 5M trigger pairs is used.

Table 6 shows that

- The MI-Trigger model is superior to word unigram and bigram models. The conditional perplexity of the DD-6-MI-Trigger model is less than that of word bigram model and much less than the word unigram model.
- The parameter number of the MI-Trigger model is much less than that of word bigram model.

One of the most powerful abilities of a person is to properly combine different knowledge. This also applies to PYCC. The word bigram model and the MI-Trigger model are merged by linear interpolation as follows:

$$\log P_{MERGED}(S) = (1-a)\cdot\log P_{Bigram}(S) + a\cdot\log P_{MI-Trigger}(S) \qquad (7)$$

where $S = w_1^n = w_1 w_2 \cdots w_n$ and a is the weight of the word bigram model. Here the DD-6-MI-Trigger model with 5M trigger pairs is applied. The result is shown in Table 7.

Table 7 shows that the recognition rate reaches up to 98.7% when the N-gram weight is 0.3 and the MI-Trigger weight is 0.7.

MI-Trigger Weight	Recognition Rate
0.0	96.2%
0.1	96.5%
0.2	97.3%
0.3	97.7%
0.4	98.2%
0.5	98.3%
0.6	98.6%
0.7	98.7%
0.8	98.5%
0.9	98.2%
1.0	97.6%

Table 7: The PYCC recognition rates of word bigram and MI-Trigger merging

Through the experiments, it has been proven that the merged model has better results over both word bigram and MI-Trigger models. Compared to the pure word bigram model, the merged model also captures the long-distance dependency of word pairs using the concept of mutual information. Compared to the MI-trigger model which only captures highly correlated word pairs, the merged model also captures poorly correlated word pairs within a short distance by using the word bigram model.

Conclusion

This paper proposes a new MI-Trigger-based modeling approach to capture the preferred relationships between words by using the concept of trigger pair. Both the distance-independent(DI) and distance-dependent(DD) MI-Trigger-based models are constructed within a window. It is found that

- The long-distance dependency is useful to language disambiguation and should be modeled properly in natural language processing.

- The DD MI-Trigger models have better performance than the DI MI-Trigger models for the same window size.
- The number of the trigger pairs in an MI-Trigger model can be kept to a reasonable size without losing too much of its modeling power.
- The MI-Trigger-based language modeling has better performance than the word bigram model while the parameter number of the MI-Trigger model is much less than that of the word bigram model. The PINYIN-to-Character conversion rate reaches up to 97.7% by using the MI-Trigger model. The recognition rate further reaches up to 98.7% by proper word bigram and MI-Trigger merging.

References

[Brent93] Brent M. "From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax". *Computational Linguistics*, Vol.19, No.2, pp.263-311, June 1993.

[Calzolori90] Calzolori N. "Acquisition of Lexical Information from a Large Textual Italian Corpus". *Proc. of COLING*. Vol.2, pp.54-59, 1990.

[Chen+87] Chen S.I. et al. "The Continuous Conversion Algorithm of Chinese Character's Phonetic Symbols to Chinese Character". *Proc. of National Computer Symposium*, Taiwan, pp.437-442. 1987.

[Chen93] Chen J.K. "A Mathematical Model for Chinese Input". *Computer Processing of Chinese & Oriental Languages*. Vol. 7, pp.75-84, 1993.

[Church90] Church K. "Word Association Norms, Mutual Information and Lexicography". *Computational Linguistics*, Vol.16, No.1, pp.22-29. 1990.

[Church+90] Church K. et al. "Enhanced Good Turing and Cat-Cal: Two New Methods for Estimating Probabilities of English Bigrams". *Computer, Speech and Language*, Vol.5, pp.19-54, 1991.

[Hindle+93] Hindle D. et al. "Structural Ambiguity and Lexical Relations". *Computational Linguistics*, Vol.19, No.1, pp.103-120, March 1993.

[Hsieh+89] Hsieh M.L. et al. " A Grammatical Approach to Convert Phonetic Symbols into Characters". *Proc. of National Computer Symposium*. Taiwan, pp.453-461, 1989.

[Hsu94] Hsu W.L. "Chinese Parsing in a Phoneme-to-Character Conversion System based on Semantic Pattern Matching". *Chinese Processing of Chinese & Oriental Languages*. Vol.8, No.2, pp.227-236, 1994.

[Kobayashi+94] Kobayashi T. et al. "Analysis of Japanese Compound Nouns using Collocational Information". *Proc. of COLING*. pp.865-970, 1994.

[Kuo96] Kuo J.J. "Phonetic-Input-to-Character Conversion System for Chinese Using Syntactic Connection Table and Semantic Distance". *Computer Processing of Chinese & Oriental Languages*. Vol.10, No.2, pp.195-210, 1996.

[Magerman+90] Magerman D. et al. "Parsing a Natural Language Using Mutual Information Statistics". *Proc. of AAAI*, pp.984-989, 1990.

[Meyer+75] Meyer D. et al. "Loci of contextual effects on visual word recognition". *In Attention and Performance V*, edited by P.Rabbitt and S.Domie. Acdemic Press, pp.98-116, 1975.

[Rosenfeld94] Rosenfeld R. "Adaptive Statistical Language Modeling: A Maximum Entropy Approach". *Ph.D. Thesis*. Carneige Mellon University, April 1994.

[Sakai+93] Sakai T. et al. "An Evaluation of Translation Algorithms and Learning Methods in Kana to Kanji Translation". *Information Processing Society of Japan*. Vol.34, No.12, pp.2489-2498, 1993.

[Shannon51] Shannon C.E. "Prediction and Entropy of Printed English". Bell Systems Technical Journal, Vol.30, pp.50-64, 1951.

[Sproat+90] Sproat R. et al. "A Statistical Method for Finding Word Boundaries in Chinese Text". *Computer Processing of Chinese & Oriental Languages*. Vol.4, No.4, pp.335-351, 1990.

[Sproat92] Sproat R. "An Application of Statistical Optimization with Dynamic Programming to Phonemic-Input-to-Character Conversion for Chinese". *Proc. of ROCLING* . Taiwan, pp.379-390, 1992.