Language, Information and Computation (PACLIC12), 18-20 Feb, 1998, 195-205

MI-Trigger-based Language Modeling

Zhou GuoDong and Lua KimTeng

Department of Information Systems and Computer Science
National University of Singapore
Lower Kent Ridge Road
Singapore 119260
{zhouguod, luakt}@iscs.nus.edu.sg

ABSTRACT

This paper proposes a new MI-Trigger-based modeling approach to capture the preferred relationships between words over a short or long distance. It is implemented by the concept of trigger pair, which is selected by average mutual information and measured by mutual information. Both the distance-independent(DI) and distance-dependent(DD) MI-Trigger-based models are constructed within a window of a size from 1 to 10. It is found that the DD MI-Trigger models have better performance than the DI MI-Trigger models for the same window size and it is better to model the preferred relationships in a distance-dependent way. 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.

Keyword: *MI-Trigger modeling approach, preferred relationship, long-distance context dependency, average mutual information, mutual information.*

1 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. One classic example is "strong" and "powerful". [Halliday66] noticed that these two words were used in different language environments such as "strong tea" and "powerful computer" although they had similar syntax and semantics. Psychologists also have a similar concept: word association. Two highly associated word pairs are "bread/butter" 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 between the two words in the 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/tea"), and solid structure(e.g. "the more/the more"). These information are useful for automatic sentence disambiguation.

The research in [Hindle+93] showed the role of correlated statistical information in resolving the sentence ambiguity. Consider a simple English sentence:

"She (wanted | placed | put) the dress on the rack."

For different verbs, the linking directions of the preposition phrase("on the rack") are different. It can modify the noun(for "wanted") or function as the objective complement of the verbs(for "placed " and "put"). This is the well-known preposition phrase(PP) attachment problem in the analysis of English. Their research showed that a reasonable analysis result can be chosen by comparing the mutual information of the verb-preposition pair("want/on") and the object-preposition pair("dress/on"). [Magerman+90] used a computational model of mutual information to automatically segment short phrases. [Rosenfeld94] used the concept of trigger pair as the basic information bearing element to extract information from further back in the document's history. Similar research includes [Brent93] and [Kobayashi+94].

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 distance to a long distance. While N-gram models are simple in language modeling and have been successfully used in speech recognition and other 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.

The organization of this chapter is as follows: Section 2 defines the concept of trigger pair. The criteria of how to select a trigger pair are described in Section 3 while Section 4 describes a method to measure the strength of a trigger pair. Section 5 describes the trigger-based language modeling approach. Section 6 gives a example of its applications: PINYIN-to-Character conversion. Finally, a summary of this paper is given in Section 7.

2 Concept of Trigger Pair

Based on the above description, we have decided to use the trigger pair 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, 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.

3 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. Let V be the size of the vocabulary. Note that, unlike the case of the bigram model where the number of different consecutive word pairs in an available corpus is always much less than V^2, the number of word pairs where both words occur in the same document is a significant fraction of V^2. Therefore, selecting a reasonable number of the most powerful trigger pairs is important to a trigger-based language model.

3.1 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 need to 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 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:

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

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. It is important that enough data should be used relative to the number of parameters in the model so that the model is not grossly over-fitted. Here, the training corpus used to compute the conditional perplexity is the automatically segmented Xin Hua news corpus, which has about 57 million characters or 29 million 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 thesis we restrict maximum window size to 10.

Distance	Conditional Perplexity	Distance	Conditional 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

3.2 Selecting Trigger Pair

We define the events w and w_o over the window:

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:

$$\log CPR(A_o, B) = \log \frac{f(A_o,B)f(\overline{A_o},\overline{B})}{f(A_o,\overline{B})f(\overline{A_o},B)} \quad (2)$$

where $f(X_o, Y)$ is the occurrence frequency 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 trigger pair consisting of two rare words(树梢 → 白皑皑), and compare it to a less well correlated, but more common trigger 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 trigger 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:

$$AMI(A_o;B) = P(A_o,B)\log\frac{P(A_oB)}{P(A_o)P(B)} + P(A_o,\overline{B})\log\frac{P(A_o\overline{B})}{P(A_o)P(\overline{B})}$$
$$+ P(\overline{A_o},B)\log\frac{P(\overline{A_o}B)}{P(\overline{A_o})P(B)} + P(\overline{A_o},\overline{B})\log\frac{P(\overline{A_o}\overline{B})}{P(\overline{A_o})P(\overline{B})} \quad (3)$$

Obviously, Equation 3 takes the joint probability into consideration. In this thesis, 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.

4 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:

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

Using Maximum Likelihood Estimation(MLE) method, we can get:

$$P(A) = \frac{f(A)}{\sum_A f(A)} \quad (5)$$

$$P(B) = \frac{f(B)}{\sum_B f(B)} \quad (6)$$

$$P(A,B) = \frac{f(A,B)}{\sum_{A,B} f(A,B)} \quad (7)$$

where $f(A)$ and $f(B)$ are the frequencies of the words A and B occurred in the corpus respectively, and $f(A,B)$ is the frequency 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 can use mutual information $MI(A_o;B)$ to measure the preference relationship degree of the trigger pair $(A_o \rightarrow B)$.

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($ws = 1,2,...,10$).

Another factor is to decide whether to measure an MI-Trigger in a distance-independent(DI) way or in a 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 Xin Hua corpus of 29 million 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 Figures 1, 2 and 3, which show 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 respectively in our Xin Hua corpus. 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).

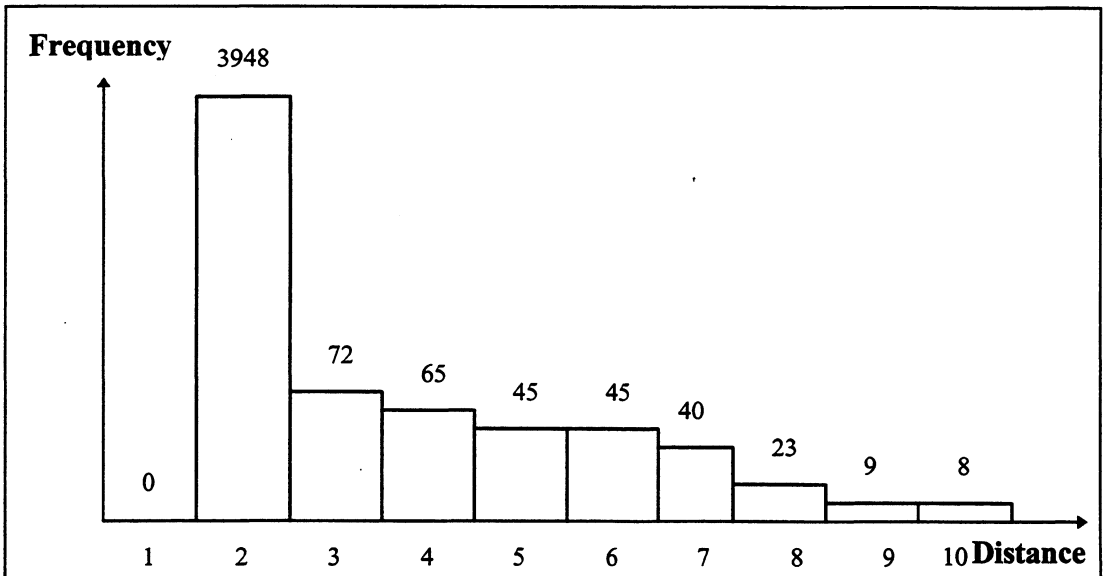

Figure 1: The occurrence frequency of the word pair("越/越") as a function of distance

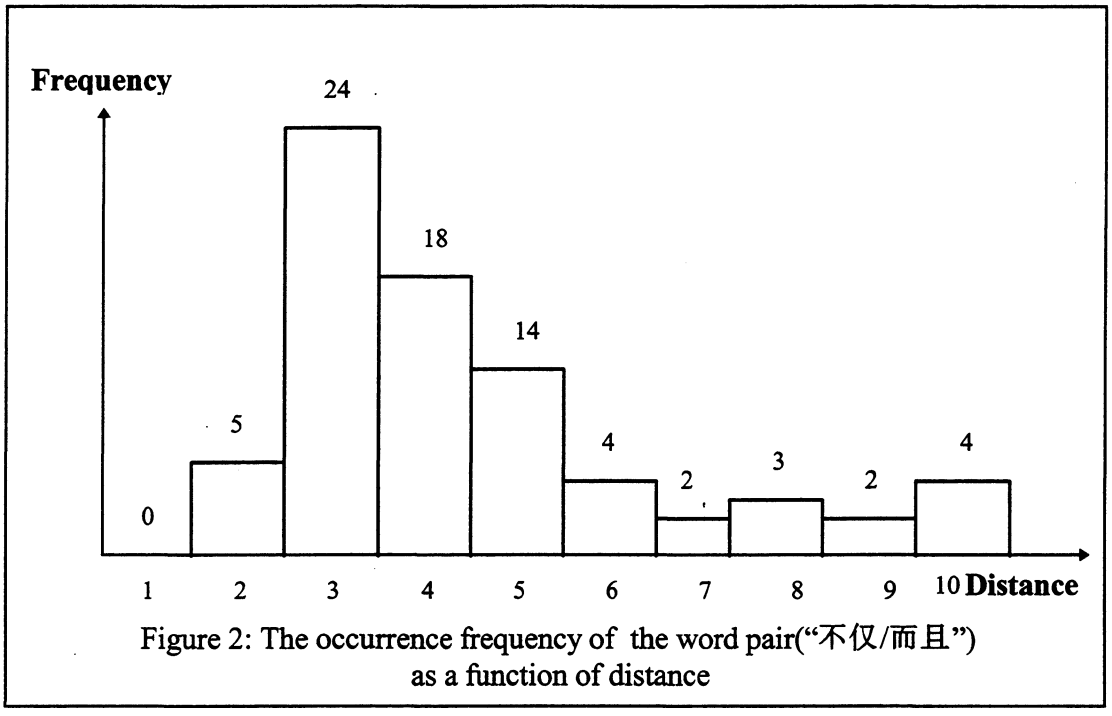

Figure 2: The occurrence frequency of the word pair("不仅/而且") as a function of distance

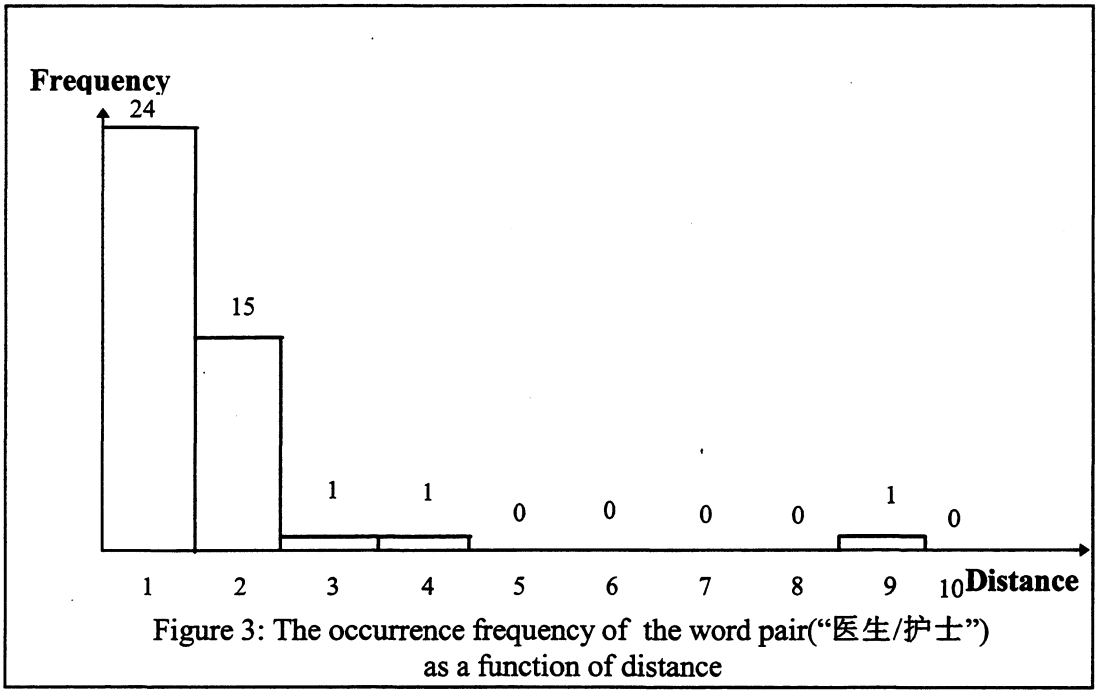

Figure 3: The occurrence frequency of the word pair("医生/护士") as a function of distance

Distance	1	2	3	4	5	6
Number of MI-Trigger Pairs (×1000)	246	219	180	146	117	92

Table 2: The numbers of the trigger pairs for different distances in the DD-6-MI-Trigger model

Average $MI_{(\times 10^{-7})}$	Number of Trigger Pairs	Average $MI_{(\times 10^{-7})}$	Number of Trigger Pairs	Average $MI_{(\times 10^{-7})}$	Number of Trigger Pairs
[1,2)	183677	[21,22)	3463	[41,42)	873
[2,3)	270035	[22,23)	3161	[42,43)	827
[3,4)	137520	[23,24)	2908	[43,44)	772
[4,5)	84219	[24,25)	2665	[44,45)	763
[5,6)	56935	[25,26)	2441	[45,46)	676
[6,7)	41025	[26,27)	2234	[46,47)	651
[7,8)	30711	[27,28)	2095	[47,48)	648
[8,9)	23517	[28,29)	1835	[48,49)	629
[9,10)	18590	[29,30)	1788	[49,50)	644
[10,11)	15240	[30,31)	1675	[50,51)	613
[11,12)	12424	[31,32)	1599	[51,52)	554
[12,13)	10482	[32,33)	1500	[52,53)	559
[13,14)	8727	[33,34)	1419	[53,54)	538
[14,15)	7710	[34,35)	1307	[54,55)	485
[15,16)	6845	[35,36)	1197	[55,56)	503
[16,17)	5946	[36,37)	1119	[56,57)	451
[17,18)	5368	[37,38)	1126	[57,58)	462
[18,19)	4644	[38,39)	972	[58,59)	432
[19,20)	4289	[39,40)	1001	[59,60)	412
[20,21)	3819	[40,41)	922	>=60	20358

Table 3: The numbers of the trigger pairs for different average mutual information in the DD-6-MI-Trigger model

MI	Number of Trigger Pairs	MI	Number of Trigger Pairs
[0,1)	9222	[10,11)	11231
[1,2)	45368	[11,12)	6814
[2,3)	300324	[12,13)	3950
[3,4)	201059	[13,14)	2116
[4,5)	154631	[14,15)	1031
[5,6)	102850	[15,16)	481
[6,7)	68055	[16,17)	190
[7,8)	45186	[17,18)	56
[8,9)	29124	[18,19)	6
[9,10)	18306	>=19	0

Table 4: The Numbers of the trigger pairs for different mutual information in the DD-6-MI-Trigger model

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 in this thesis. Each model differs in different window sizes, and whether the evaluation is done in the DI way or in the DD way. Moreover, for ease of comparison, each MI-Trigger model includes the same number of the best trigger pairs while the values of all the other word pairs are set to 0. In our experiments, only the best 1 million 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 6.

For simplicity in this paper, we can 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 Xin Hua corpus of 29 million words. Let' s take the DD-6-MI-Trigger model as a example. We filter about $28 \times 28 \times 6$ million(with six different distances and with about 28000 Chinese words in the lexicon) possible distance-dependent word pairs. As a first step, only word pairs that co-occur at least 3 times are kept. This results in 5.7 million word pairs. Then selected by average mutual information, the best 1 million word pairs are kept as trigger pairs. Finally, each of the best 1 million MI-Trigger pairs are

measured by mutual information. In this way, we build a DD-6-MI-Trigger model which includes the best 1 million trigger pairs which are measured by mutual information. Some of statistics about the DD-6-MI-Trigger model are shown in Table 2, Table 3 and Table 4. These tables show the number of trigger pairs for different distances, average mutual information and mutual information respectively.

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 $\log P(S)$ as follows:
For a distance-independent MI-Trigger-based model,

$$\log P(S) = \sum_{i=1}^{n} \log P(w_i) + \sum_{i=n}^{2} \sum_{j=i-1}^{\max(1,i-ws)} DI - ws - MI - Trigger(w_j \rightarrow w_i) \quad (8)$$

and for a distance-dependent MI-Trigger-based model,

$$\log P(S) = \sum_{i=1}^{n} \log P(w_i) + \sum_{i=n}^{2} \sum_{j=i-1}^{\max(1,i-ws)} DD - ws - MI - Trigger(w_j \rightarrow w_i, i-j+1) \quad (9)$$

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 8 and 9 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.

6 PINYIN-to-Character conversion

As an application of the MI-Trigger modeling approach, a PINYIN-to-Character conversion system is constructed. In fact, PINYIN-to-Character conversion has become one of the basic problems in Chinese language processing and has been the subjects of many researchers in the last decade. There are several kinds of approaches used in such tasks, including:

- The longest word preference algorithm[Kuo+86][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 including lexicon, syntactic and semantic rules[Hsieh+89][Yeh+91][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[Chang+91][Sproat92][Chen93][Lin87]. 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 application of PINYIN-to-Character conversion. For ease of comparison, the PINYIN counterparts of 600 Chinese sentences(6104 Chinese characters) from Singapore primary school Chinese text books are used for testing.

The PINYIN-to-Character conversion rates of different MI-Trigger models are shown in Table 5.

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 5: The PINYIN-to-Character conversion rates of the MI-Trigger models

It is found from Table 5 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 distance-dependent way. It is also found that the PINYIN-to-Character conversion rate can reach up to 96.6%.

As it was stated above, all the MI-Trigger models only include the best 1 million 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. The PINYIN-to-Character conversion rates are shown in Table 6.

Number of the MI-Trigger Pairs in the DD-6-MI-Trigger Model	Recognition Rate
0	85.3%
100,000	90.7%
200,000	92.6%
400,000	94.2%
600,000	95.5%
800,000	96.3%
1,000,000	96.6%
1,200,000	96.7%
1,400,000	96.8%
1,600,000	97.0%
1,800,000	97.1%
2,000,000	97.2%
2,500,000	97.1%
3,000,000	97.2%
3,500,000	97.2%
4,000,000	97.3%
4,500,000	97.4%
5,000,000	97.6%
6,000,000	97.7%

Table 6: The effect of different numbers of the trigger pairs in the DD-6-MI-Trigger model on the PINYIN-to-Character conversion rates

We can see from Table 6 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.

7 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. The trigger pairs are selected by average mutual information and measured by mutual information. Both the distance-independent(DI) and distance-dependent(DD) MI-Trigger-based models are constructed within a window of a size from 1 to 10. It is found that

- The DD MI-Trigger models have better performance than the DI MI-Trigger models for the same window size. Therefore, it is better to model the preferred relationships between words in a distance-dependent way.
- 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 long-distance dependency is useful to language disambiguation and should be modeled properly in natural language processing.

8 Reference

[Brent93] Brent M. "From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax". *Computational Linguistics*, Volume 19, Number 2, pp. 263-311, June 1993

[Chang+91] Chang J.S. et al. "Conversion of Phonetic-Input to Chinese Text through Constraint Satisfaction". *Proceedings of ROCLING IV*. Taiwan, pp. 30-36, 1991.

[Chen+87] Chen S.I. et al. "The Continuous Conversion Algorithm of Chinese Character's Phonetic Symbols to Chinese Character". *Proceedings of 1987 National Computer Symposium*, Taipei, 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.

[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*, Volume 5, pages 19-54, 1991.

[Halliday66] Halliday M. "Lexis as a linguistic level". *In memory of J.R.Firth*, edited by C.Bazell, J.Catford, M.Halliday and R.Robins, Longman, 1966.

[Hindle+93] Hindle D. et al. "Structural Ambiguity and Lexical Relations". *Computational Linguistics*, Volume 19, Number 1, Pages 103-120, March 1993.

[Hsieh+89] Hsieh M.L. et al. " A Grammatical Approach to Convert Phonetic Symbols into Characters". *Proceedings of 1989 National Computer Symposium*. Taipei, 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". *Proceedings of COLING*. pp.865-970, 1994.

[Kuo+86] Kuo J.J. et al. "The Development of New Chinese Input Method-Chinese-Word-String Input System". Proceedings of 1986 International Computer Symposium, Taiwan, pp. 1470-1479, 1986.

[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.

[Lin87] Lin M.Y. "Removing the Ambiguity of Phonetic Chinese Input by Relaxation Technique". *Computer Processing of Chinese & Oriental Languages*. Vol. 3, pp. 1-24, 1987.

[Magerman+90] Magerman D. et al. "Parsing a Natural Language Using Mutual Information Statistics". *Proceedings of AAAI-90*, 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.Dornie. Acdemic Press, 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.

[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". *Proceedings of ROCLING V*. Taiwan, pp. 379-390, 1992.

[Yeh+91] Yeh C.L. et al. "Rule-based Word Identification for Mandarin Chinese Sentences-A Unification Approach". *Computer Processing of Chinese & Oriental Languages*. Vol. 5, No. 2, pp.97-118, 1991.