Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008

Disambiguating Effectively Chinese Polyphonic Ambiguity Based on Unify Approach

Feng-Long Huang

Department of Computer Science and Information Engineering
National United University
No. 1, Lienda, Miaoli, Taiwan, R. O. C. 36003
flhuang@nuu.edu.tw

Abstract

One of the difficult tasks on Natural Language Processing (NLP) is to resolve the sense ambiguity of characters or words on text, such as polyphones, homonymy, and homograph. The paper addresses the ambiguity issue of Chinese character polyphones and disambiguity approaches for such issues. Three methods, dictionary matching, language models and voting scheme, are used to disambiguate the prediction of polyphones. The best precision rate for these methods achieves 92.65%. Furthermore we proposed the unify approaches to improve the performance with respect to various threshold value. Comparing with the well-known MS Word 2007, our approach is superior and enhances the final precision rate up to 93.32%.

Keywords: Sense Disambiguity, Language Model, Voting Scheme, Unify Approach.

1. Introduction

In recent years, natural language processing (NLP) has been studied and discussed on many fields, such as machine translation, speech processing, lexical analysis, information retrieval, spelling prediction, hand-writing recognition, and so on [1][2]. In the computational models, syntax models parsing, word segmentation and generation of statistical language models have been the focus tasks.

In general, no matter what kinds of natural languages, there will be always a phenomenon of ambiguity among characters or words in sentences, such as polyphone, homonymy, homograph, and the combination of them. The issues are so-called word sense dsiambiguity (WSD)[3][4]. One of the difficult tasks on NLP is to resolve the sense ambiguity.

Disambiguating the issues of sense ambiguity can alleviate the problems in NLP. The paper address the dictionary matching, statistical *N*-gram language model and voting scheme, which includes two methods: preference and winner-take-all scoring, to retrieve the Chinese lexical knowledge in sentence, employed to process WSD on Chinese polyphonic characters. There are near 5700 frequent unique characters and among them more than 1300 characters have more than 2 different pronunciations, they are called polyphonic characters.

The paper is organized as following: the related works on WSD are presented in Section 2. Three methods will first be described in Section 3 and experimental results are shown and then analyzed furthermore in Section 4. The proposed unify approaches are described in detail in section 5 and the conclusions and future works are listed in last section.

2. Related works

Resolving automatically the word sense ambiguity can enhance the language understanding, which will used on several fields, such as information retrieval, document category, grammar analysis, speech processing and text preprocessing, and so on. In the past decades, ambiguity issues are always considered as AI-complete. Based on the generation of large amount of machine readable text, WSD has been one of important tasks on NLP.

The approaches [6] on WSD are categorized as follows:

1) **Machine-Readable Dictionaries (MRD):**
   Relying on the word information in dictionary for ambiguity.
2) **Computational Lexicons:**
   Employing the lexical information, such as the well-known *WordNet* [7], which contains the lexical clues of characters and lattice among related characters
3) **Corpus-based methods**
   Such as part-of-speech (POS), frequency and location of characters and words[8].

There are many works addressing WSD and several methods have been proposed so far. Because of the unique features of Chinese language, such as Chinese word segmentation, more than two different features will be used to achieve higher prediction rate. Therefore, two methods will be arranged furthermore.

3. Description of Proposed Methods

In this paper, several methods are proposed to disambiguate the polyphones of Chinese characters; Dictionary Matching, Language Models and voting Scheme

3.1 Dictionary Matching

In order to predict correctly the pronunciation category of polyphones, dictionary matching will be exploited for the ambiguity issue. Within a Chinese sentence, the location of polyphonic character C_p is set as the centre, we extract the right and left substring based on the centre C_p. Two substrings are denoted as CH_L and CH_R. In a window size, all possible substrings in CH_L and CH_R will be segmented and then match the lexicons in dictionary.

If the words are existed on both substrings, then we can decide the pronunciation of polyphone based on the priority of longest word and highest frequency of word; length of word first and then frequency of word secondly. In the paper, window size=6 Chinese characters; that means LEN(CH_L)= LEN(CH_R)=6。

The Chinese dictionary is available and contains near

978-1-4244-2096-4/08/$25.00 ©2008 IEEE

Authorized licensed use limited to: BEIJING UNIVERSITY OF TECHNOLOGY. Downloaded on June 22,2022 at 03:00:59 UTC from IEEE Xplore. Restrictions apply.

130K Chinese words (zhong1 wen2 ci2, 中文詞). Each Chinese word may be composed from 2 to 12 Chinese characters (zhong1 wen2 zi4, 中文字). All the words in dictionary contain its frequency, POS, and pronunciation (Juu4 yin1 fu2 hau4, 注音符號); which decided correctly pronunciation of polyphonic character in the word.

The algorithm of dictionary matching is described as:

step 1. Read in the sentence and find the location C_p of polyphone.
step 2. Based on the of C_p, all the possible substring of CH_L and CH_R within window (size=6) will be segmented and extracted, then compared with lexicons in Chinese dictionary.
step 3. If any Chinese word can be found on both substring
goto step 4,
else
goto step 5.
step 4. Decide the pronunciation of polyphone on the priority of longest word and then the highest frequency of word. The process ends.
step 5. The pronunciation of polyphone C_p will be predicted by sequential methods.

3.2 Language Models - LMs

In recent years, the statistical language models have been used in NLP. Supoosed that $W=w_1,w_2,w_3,\ldots w_n$, where w_i and n denote the the i^{th} Chinese character and its number in a sentence $(0 \leq i \leq n)$。

$P(W)=P(w_1, w_2 \ldots, w_n)$, //using chain rules.

$$P(w_1^n)= P(w_1)P(w_2|w_1)P(w_3|w_1^2)\ldots P(w_n|w_1^{n-1})$$
$$=\prod_{k=1}^{n} P\left(w_k \middle| w_1^{k-1}\right) \tag{1}$$

where w_1^{k-1} denotes string $w_1,w_2,w_3,\ldots w_{k-1}$.

In Eq(1), we calculate the probability $P\left(w_k \middle| w_1^{k-1}\right)$, starting at w_1, by using $w_1,w_2,w_3\ldots w_{k-1}$ substring to predict the occurrence probability of w_k. In case of longer string, it is necessary for large amount of corpus to train the language model with better performance. It will lead to spending much labor and time extensive.

In general, unigram, bigram and trigram (3<=N) [5] are generated. N-gram model calculates probability $P(\cdot)$ of N^{th} events by the preceding N-1 events, rather than string $w_1,w_2,w_3\ldots w_{N-1}$.

In short, N-gram is so-called N-1)th-order Markov model, which calculate conditional probability of successive events: calculate the probability of N^{th} event while preceding (N-1) event occurs. Basically, N-gram Language Model is expressed as follows:

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-N+1}^{k-1}) \tag{2}$$

N=1, unigram or zero-order markov model.
N=2, bigram or first-order markov model.
N=3, trigram or second-order markov model.

In Eq(2), the relative frequency will be used for calculating the $P(\cdot)$:

$$P(w_n|w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1} w_n)}{C(w_{n-N+1}^{n-1})}, \tag{3}$$

where $C(w)$ denotes the count of event w occurring in training corpus.

In Eq(3), the obtained probability $P(\cdot)$ is called Maximum Likelihood Estimation (MLE). While predicting the pronunciation category of polyphones, we can predict based on the probability on each category t ($1 \leq t \leq T$), T denotes the number of categories of polyphone. The category with maximum probability $P_{max}(W)$ with respect to the sentence W will be the target and then the correct pronunciation of polyphone can be decided.

3.3 Voting Scheme

In contrast to the N-gram models above, we proposed voting scheme with similar concept for use to select in human being society. Basically, we vote for one candidate and the candidates with maximum votes will be the winner. In real world, maybe more than one candidate will win the section game while disambiguation process only one category of polyphone will be the final target with respect to the pronunciation.

The voting scheme can be described as follows: each token in sentence play the voter for vote for favorite candidate based on the probability calculated by the lexical features of tokens. The total score $S(W)$ accumulated from all voters for each category will be obtained, and the candidate category with highest score is the final winner. In the paper, there are two voting methods:

1) Winner-Take-All:

In the voting method, the probability is calculated as follows:

$$P(w_i) = \frac{C(w_i,t)}{C(w_i)} \tag{4}$$

where $C(\mathrm{w_i})$ denotes the ouucrrences of w_i in training corpus, and $C(w_i, t)$ denotes the occurrences of token w_i on category t.

In Eq(4) above, $P(w_i)$ is regarded as the probability of w_i on category t. In winner take all scoring, the category with maximum probability will win the ticket. On the other hand, it win one ticket (1 score) while all other categories can't be assigned any ticket (0 score). Therefore, each voter has just one ticket for voting. The winner-take-all scoring for tolen w_i is defined as follows:

$$P(w_i) = \begin{cases} 1 & \text{if } P_t(w_i) => \text{max} \\ 0 & \text{all other categories} \end{cases} \tag{5}$$

Based on the Eq(5), the total score for each categories can be accumulated for all tokens in sentence:

$$S(W) = P(w_1)+P(w_2)+P(w_3)+\ldots+P(w_n)$$
$$=\sum_{k=1}^{n} P(w_k) \tag{6}$$

2) Preference Scoring:

Another voting method is called as preference. For a token in sentence, the summation of the probability for all the categories of a polyphone character will be equal to 1. Let us show an example (E1) for two voting methods. As presented in Table 1, the polyphone character 卷 has three different pronunciations, 1 ㄐㄩㄢˋ, 2 ㄐㄩㄢˇ and 3 ㄑㄩㄢˊ. Supposed that the occurrence of token 白卷 (blank examination) in these categories are 26, 11 and 3, total occurrence is 40. Therefore, the score for each category by two scoring methods can be calculated.

教育社會方面都繳了白卷 (E1)
Government handed over a blank examination paper in education and society.

Authorized licensed use limited to: BEIJING UNIVERSITY OF TECHNOLOGY. Downloaded on June 22,2022 at 03:00:59 UTC from IEEE Xplore. Restrictions apply.

Table 1: example of two scoring scheme.

category	count	preference	w-t-all
1 ㄐㄩㄢˋ	26	26/40=0.65	40/40=1
2 ㄐㄩㄢˇ	11	11/40=0.275	0/40=0
3 ㄑㄩㄢˊ	3	3/40=0.075	0/40=0
Total ∑	40	1 score	1 score

ps. w-t-all denotes winner-take-all scoring

3.4 Unknown events-Zero count

In certain cases, $C(\cdot)$ of a novel, which don't occur in the training corpus, may be zero because of the limited training data and infinite language. It is always hard for us to collect sufficient datum. The potential issue of MLE is the probability for unseen events is exactly zero. This is so-called the zero-count problem and will degrade the performance of system.

It is obvious that zero count will lead to the zero probability of $P(\cdot)$ in Eqs(2), (3) and (4). The paper adopted the additive discounting for calculating P^* as follows:

$$p^* = (c+\delta)\frac{N}{N+B\delta}. \quad (7)$$

where δ denotes a small value (δ<=0.5); which will be added into all the known and unknown events. The smoothing method will alleviate the zero count issue in language model.

3.5 Classifier-Predicting the Categories

Supposed that polyphone has T categories, $1 \le t \le T$, how can we predict the correct target $\hat{t}$? As shown in Eq(8), the category with maximum probability or score will be the most possible target:

$\hat{t}=argmax_t P_t(W)$, or

$$\hat{t}=argmax_t S_t(W), \quad (8)$$

where $P_t(W)$ is the probability of W in category t, which can be obtained from Eq(1) for LMs and $S_t(W)$ is the total score based on the voting scheme from Eq(6).

4. Experiment Results

In the paper, 10 Chinese polyphones are selected randomly from more than 1300 polyphones in Chinese. In the following, we first introduce the dictionary and corpus adopted in the paper.

4.1 Dictionary and Corpus

Academic Sinica Chinese Electronic dictionary, ASCED) contains more than 130K Chinese words, composing of 2 to 11 characters. The word in ASCED is with Part-of-speech (POS), frequency and pronunciation for each character.

The experimental data are collected from the corpus of Sinica and news from China Times. The sentences with one of 10 polyphone characters are collected randomly. There are totally 9070 sentences, which are divided into two parts: 8030 (88.5%) and 1040 (11.5%) sentences for training and outside testing, respectively.

4.2 Experiment Results

Three LMs models are generated: unigram, bigram and trigram. Precision Rate (PR) can be defined as:

$$\text{PR} = \frac{\text{NO. of correct prediction}}{\text{total number of sentence}} \quad (9)$$

Method 1: Dictionary Matching

The predicted results are shown in Table 2. There are 69 sentences processed by the matching phase and 7 sentences are wrongly predicted. The average PR achieves 89.86%.

In the followings, several examples are presented and explained the matching phase of dictionary:

我們回頭看看中國人的歷史。 (E2)

We look back the history of Chinese.

Based on the matching algorithm, two substring CH_L and CH_R of polyphone 中 for (E2);

CH_L =”們回頭看看中”,

CH_R=” 中國人的歷史”.

Upon the word segmentation, the Chinese word and pronunciation are as follows:

CH_L			CH_R		
看中	83	ㄓㄨㄥˋ	中國	3542	ㄓㄨㄥ
			中國人	487	ㄓㄨㄥ

According the priority of length of word first,中國人 (Chinese people) will decide the pronunciation of 中 as ㄓㄨㄥ.

看中文再用廣東話來發音。 (E3)

Read the Chinese and then pronounce in Canton.

Chinese words in CH_L			Chinese words in CH_R		
看中	83	ㄓㄨㄥˋ	中文	343	ㄓㄨㄥ

峰迴路轉再看中國方面. (E4)

The path winds along mountain ridges, then watch the reflection of China.

Chinese words in CH_L			Chinese words in CH_R		
看中	83	ㄓㄨㄥˋ	中國	3542	ㄓㄨㄥ

中央研究院未來的展望。 (E5)

The future forecast of Academic Sinica of Chinese.

Chinese words in CH_L	CH_R		
	中央	2979	ㄓㄨㄥ
	中央研究院	50	ㄓㄨㄥ

In example (E5), only CH_R contains the segmented words. On the other hand, there are no word in CH_L

Method 2: Language Model (LMs)

The experiment results of three models unigram,

Authorized licensed use limited to: BEIJING UNIVERSITY OF TECHNOLOGY. Downloaded on June 22,2022 at 03:00:59 UTC from IEEE Xplore. Restrictions apply.

bigram, trigram are listed in Table 3. Bigram LMs achieves 92.58%, which is highest rate among three models.

Method 3: Voting Scheme

1)Winner take all: Three models; unitoken, betoken and tritoken are generated. As shown in Table 4. Bitoken achieves highest PR of 90.17%.

2)Preference: Three models; unitoken, bitoken and tritoken are generated. As shown in Table 5. Bitoken preference scoring can achieves highest PR of 92.72% in average.

4.3 Word 2007 precision rate

MS Office is a famous and well-known editing package around world. In our experiments, MS Word 2007 is used to process the transcription on same testing sentences. PR achieves 89.8% in average, as shown in Table 6.

4.4 Results Analysis

In the paper, voting scheme of preference and winner-take-all scoring, and statistical language Model have been proposed and employed to resolve the issue of polyphone ambiguity. We compare these methods with MS Word 2007. Preference bitoken achieves highest PR among these models and achieves 92.72%. It is apparent that our proposed methods are all superior to MS Word 2007.

In the following, two examples are shown for correct and wrong prediction by Word 2007.

ㄐㄧㄠˋ ㄩˋ ㄕㄜˋ ㄏㄨㄟˋ ㄈㄤ ㄇㄧㄢˋ ㄉㄡ ㄐㄧㄠˇ ˙ㄌㄜ ㄅㄞˊ ㄐㄩㄢˋ
教 育 社 會 方 面 都 繳 了 白 卷

Government handed over a blank examination paper in education and society. (correct prediction)

ㄅㄤ ㄖㄨㄛˋ ㄨˊ ㄖㄣˊ ㄅㄢ ㄗˋ ㄧㄢˊ ㄗˋ ㄩˇ
傍 若 無 人 般 自 言 自 語

Talking to oneself as if nobody is around.(wrong prediction)

5. Unify Approach - Further Improvement

Upon all our experiments above, the best precision rate obtained from bitoken Preference scoring in model 3 achieves 92.72%.The 2^{nd} and 3^{rd} higher PR are 92.58% and 90.1% obtained from bitoken winner take all in model 3 and bigram language model, respectively. How can we enhance furthermore the prediction for the Chinese polyphonic ambiguity?

5.1 The Alternative Methods

If the scores difference of top 2 predicted categories is so close or 0, it means the feature of used by current method may get higher polyphonic ambiguity and lead to wrong prediction. Various methods employ the different feature of Chinese syntax and semantic information in sentence. In the paper, the unify approach has been proposed to resolve the issue effectively. Several methods are unified effectively to resolve the prediction of polyphones based on alternative algorithm.

Basically, the best method of Preference scoring is employed to predict the category, based on the voting scheme. While the difference of score for top 2 predicted categories is equal or less than a threshold value, in such situation, the alternative method with different feature will be triggered and play as the alternate method to predict.

5.2 Unify approach

As shown in Figure 1, unify approach merges bitoken preference scoring with the alternative-bigram language model. The threshold value θ is set to decide whether the alternative will be triggered or not. While the difference of top 2 scores calculated by the bitoken preference scoring in model 3 is less than or equal to θ, the alternative-bigram language model (LMs) will be triggered.

It is apparent that the threshold will affect the number of sentences processed by the alternative and final precision rate. In the paper, we implemented and evaluated the various θ and then got the best value to achieve best performance.

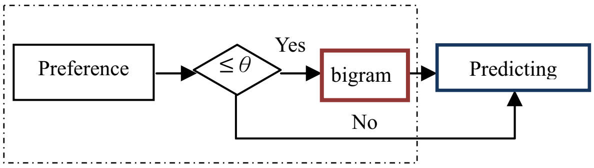

Figure 1: Unify approach: bigram LM is the alternative

In our experiments, we search various θ to find better results. There are 26 sentences predicted by the bigram under $\theta = 0.01$ and 8 sentences predicted wrong by bitoken preference scoring are correctly predicted by the alternative bigram LMs, as presented in Table 7.

The unify approach improves 0.6% (93.32-92.72) higher in average and final PR achieves 93.32%. It proves that our approach outperforms MS Word.

6. Conclusion

In the paper, we used several methods to resolve the issue of ambiguity of Chinese polyphones. First, three methods are employed to predict the category of polyphone: dictionary matching, language models and voting scheme; the last method has two different scoring methods: winner-take-all and preference scoring. Furthermore we propose the unify approaches to improve the prediction based on the alternative algorithm and. the best result is 93.32%. The adopted approach in the paper outperforms MS Word and net precision rate is enhanced up to 3.52% (93.32-89.8).

In future, several researches will be studied:

- Collecting more corpus and extend the proposed methods to other Chinese polyphones.
- More lexical features, such as location and semantic information, used to enhance the precision rate of prediction.
- Improving the smoothing techniques for unknown words.
- Bilingual translation cross English and Chinese.

Acknowledgement

The paper is supported partially under the Project: 96-2815-C-239-022–E, NCS, Taiwan, R. O. C. Author world like to thank Dr. M. S. Yu for his valuable suggestion.

Authorized licensed use limited to: BEIJING UNIVERSITY OF TECHNOLOGY. Downloaded on June 22,2022 at 03:00:59 UTC from IEEE Xplore. Restrictions apply.

Reference

[1] Yan Wu, Xiukun Li and Caesar Lun, 2006, A Structural-Based Approach to Cantonese-English Machine Translation, Computational Linguistics and Chinese Language Processing, Vol. 11, No. 2, June 2006, pp. 137-158

[2] Brian D. Davison, Marc Najork, Tim Converse, 2006, SIGIR Workshop Report, Vol. 40 No. 2.

[3] Oliveira, F.; Wong, F.; Li, Y.-P. ,2005, Machine Learning and Cybernetics,. Proceedings of 2005 International Conference on Volume 6, Issue , 18-21 Aug. 2005 Vol. 6, An unsupervised & statistical word sense tagging using bilingual sources, Page(s): 3749 - 3754

[4] E. Agirre, P. Edmonds, 2006, Word Sense Disambiguation Algorithms and Applications, Springer.

[5] Jurafsky D. and Martin J. H., 2000, Speech and Language Processing, Prentice Hall.

[6] Nancy Ide and Jean Véronis, 1998, **24**(1) , Word Sense Disambiguation,The state of the art Computational Linguis- tics, pp. 1~41.

[7] Miller, George A.; Beckwith, Richard T. Fellbaum,Christiane D.; Gross, Derek; and Miller, Katherine J. (1990). WordNet: A non-line lexical database. International Journal of Lexicography, **3**(4), 235-244.

[8] Church, Kenneth W. and Mercer, Robert L.(1993). Introduction to the Special Issue on Computational Linguistics using Large Corpora. Computational Linguistics, **19**(1), 1-24.

Table 2: dictionary matching; total processed sentences of outside testing and precision rate (PR).

	中	乘	乾	了	傍	作	著	卷	咽	從	**total/avg.**
total sentences	33	4	0	0	0	16	11	1	0	4	69
error no.	2	0	0	0	0	1	3	1	0	0	7
Precision rate	93.94	100	--	--	--	93.75	72.73	0	--	100	89.86

Table 3：PR of outside testing on Language Model.

	中	乘	乾	了	傍	作	著	卷	咽	從	**avg.**
unigram	95.88	86.84	92.31	70.21	85.71	96.23	75.32	100	98	91.67	89.98
bigram	96.75	84.21	96.15	85.11	92.86	94.34	81.17	96.30	100	93.52	**92.58***
trigram	80.04	57.89	61.54	58.51	78.57	52.83	60.39	62.96	88	71.30	70.50

Table 4: PR of outside testing on Winner-take-all scoring.

	中	乘	乾	了	傍	作	著	卷	咽	從	**avg.**
unitoken	96.96	84.21	80.77	57.45	71.43	94.34	58.44	85.19	84	87.04	84.69
bitoken	96.75	86.84	96.15	79.79	85.71	92.45	68.83	100	98	93.52	**90.17***
tritoken	79.83	60.53	61.54	60.64	78.57	52.83	59.74	66.67	88	71.3	70.69

Table 5: PR of outside testing on Preference scoring.

token	中	乘	乾	了	傍	作	著	卷	咽	從	**avg.**
unitoken	96.96	84.21	80.77	70.21	71.43	94.34	70.13	85.19	88	87.96	87.76
bitoken.	96.75	86.84	96.15	87.23	85.71	93.40	81.17	100	98	93.52	**92.72***
tritoken.	80.04	60.53	61.54	60.64	78.57	52.83	59.74	66.67	88	71.30	70.78

Table 6: PR of MS Word 2007 on same testing sentences.

	中	乘	乾	了	傍	作	著	卷	咽	從	**avg.**
word 2007	93.37	76.47	76.67	83.65	78.57	93.70	78.33	82.76	100	91.51	**89.80**

Table 7: Unify approach, sentences predicted by the alternative under θ=0.01.

A→B	中	乘	乾	了	傍	作	著	卷	咽	從	total
T→F	0	0	0	0	0	0	0	0	0	0	0
F→T	0	0	0	1	1	1	3	0	1	1	8
F→F	2	2	0	1	1	4	6	0	0	1	17
T→T	0	0	0	0	0	0	1	0	0	0	1

ps: A→B：bitoken preference scoring→bigram Language Model, T: correct prediction, F: wrong prediction.

Authorized licensed use limited to: BEIJING UNIVERSITY OF TECHNOLOGY. Downloaded on June 22,2022 at 03:00:59 UTC from IEEE Xplore. Restrictions apply.