Disambiguation for Polyphones of Chinese Based on Two-Pass Unified Approach

Feng-Long Huang, Jun-Hong Lin, Xin-Wei Lin

Department of Computer Science and Information Engineering, National United University

No. 1, Lienda, Miaoli, Taiwan, R. O. C. 36003
flhuang@nuu.edu.tw

Abstract

The paper addresses the issue of Chinese polyphones and the disambiguity approach. Three methods, dictionary matching, language models and voting scheme, are used to retrieve the token's information and then disambiguate the prediction of polyphones. The best precision for these methods achieves 92.72%. Furthermore we proposed the two-pass unified approach to improve the performance with various empirical thresholds. Our approach is superior to the well-known MS Word 2007, and the final precision rate reaches 94.3%. It proves that the proposed approach can resolve effectively text-to-phoneme conversion in Chinese TTS system.

Keywords: Word Sense Disambiguation, Language Model, Two-Pass Unified Approach, Text-to-Speech(TTS).

1. Introduction

With the continuous development of information technology and Internet, by the digital learning approaches, learners can create a self leaning environment without physical limitation. The development of computer and network has leaded us into e-leaning era. E-learning systems have been widely adopted by numerous organizations and schools in recent years. It has also been proved to be effective and efficient to improve learners' knowledge. Several works are found in [1][2][3][7].

Chinese Language has several unique characteristics, such as the speech of transcription and sentence of characters without space between them. There are 1317 Chinese polyphones which have more than one pronunciation with different meanings. The words with such polyphones always possess multiple meanings; which can be one kind of sense ambiguity of language. It is a difficult task for all the learners, not only elementary and junior school students but also the generals, to understand the knowledge for Chinese polyphones of pronunciation with different meanings. In addition, resolving the translation and transcription of simplified, traditional Chinese, Tongyong Pinyin, Hanyu Pinyi, Taiwan Phonetics and Mandarin phonics symbol II (MPS II) will be very helpful for the generals, including all the students, to learn Chinese.

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 [4][5]. Resolving the categories of polyphones in text analysis process on Text-to-speech (TTS) system 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 then the combination of them. The issues are so-called sense ambiguity. Disambiguating the issues of word sense disambiguation (WSD) can alleviate the problems of ambiguity in NLP. Therefore, disambiguating the Chinese polyphones is the issue of text-to- phoneme conversion.

The paper address the dictionary matching, *N*-gram language model and voting scheme, which includes two scoring methods: preference and winner take all, to retrieve the Chinese lexical knowledge. The lexical information will be employed to process WSD on Chinese polyphonic characters. There are near 5700 frequent unique characters and among them 1300 characters have more than 2 different phone categories, they are so-called the polyphonic characters.

There are many works addressing approaches to exploit the semantic or syntax features that characterize learning objects and learner profile, developed an active learning system with semantic support for learners to access and navigate through learning resources in an efficient and personalized manner[6][7] [10].

There are many works [9][10][11] can be found for WSD. 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. As shown in Table 1, the Chinese polyphone "和", five phones and related senses are presented.

Due to the size limitation of paper, we focus on the approach of WSD for the Chinese polyphones while our purposed online system can translates the transcription of simplified, traditional Chinese and Pinyi.

In the following, the paper is organized: Several proposed methods will be presented in Section II. Experimental results are shown and compared with Word 2007 in section III. Furthermore, the improving approach-two-pass unified approach will be described in section IV. The conclusions and future topics are listed in last section.

978-1-4244-7640-4/10/$26.00 ©2010 IEEE

Table 1: the polyphone "和", 5 phones and related senses

category	Pinyin[1]	Sense	Chinese example
1	he2	harmonious	和諧 hexie
2	han4	and	我和你 wohanni
3	huo4	to mix dough	和麵 huomian
4	he4	to repeat what others say	唱和 changhe
5	huo5	warm	暖和 nuanhuo

2. Proposed Methods

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

2.1 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 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.

2.2 Language Models - LMs

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

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

$P(w_1^n)= P(w_1)P(w_2|w_1)P(w_3|w_1^2)...P(w_n|w_1^{n-1})$

$$=\prod_{k=1}^{n} P(w_k|w_1^{k-1}) \quad (1)$$

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

In formula (1), we calculate the probability $P(w_k|w_1^{k-1})$, starting from w_1, by using $w_1,w_2,w_3...w_{k-1}$ 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) [are generated. N-gram model calculates $P(\cdot)$ of N^{th} events by the preceding N-1 events, rather than string $w_1,w_2,w_3...w_{N-1}$.

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:

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

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

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

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

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

2.3 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.

[1] number 1,2,3,4,5 in **Pinyin** denote the Chinese tone 1,2,3,4 and softtone, respectively.

In the paper, there are two voting methods:

1) Winner Take All:

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

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

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

In formula (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 voting scheme is as follows:

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

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

$$S(W) = P(w_1)+P(w_2)+P(w_3)+\ldots+P(w_n)$$

$$=\sum_{k=1}^{n} P(w_k) \quad (6)$$

2) Preference Scoring:

Another voting method is called as preference scoring. For a token in sentence, the probability assigned to each sense will be calculated based on the relative frequency of token on such sense. The summation of the probability for all the categories of a polyphone "**和**" will be equal to 1. Let us show an example (E1) for two voting methods.

你 和 我 都 是 大 學 生 (E1)
ni han wo dou shi da xue sheng
You and I are all colleges students.

2.4 Unknown events-Zero count

As shown in Eq. (4), *C*(・) 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 that the probability for unseen events is exactly zero. This is so-called the zero-count problem. It is obvious that zero count will lead to the zero probability of *P*(・) in Eqs. (3) and (4).

There are many smoothing works in [6][7]. 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.

2.5 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 formula (8), the category with maximum probability or score will be the most possible target:

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

$\hat{t}=argmax_t S_t\ (W)$, (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 Eqs(6).

3. Experiment Results

In the paper, 42 Chinese polyphones are selected randomly as predicted targets from more than 1000 polyphones in Chinese language.

3.1 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. 40 polyphonic characters are selected randomly from more than 1000 Chinese polyphones. There are totally 36,500 sentences, which are divided into two parts: 31390 (88.6%) and 5110 (11.4%) sentences for training and testing respectively.

3.2 Experiment Results

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

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

Method 1: Dictionary Matching

There are 275 sentences processed by the matching phase and 29 sentences are predicted wrong. The PR reaches 89.44%. In the following, several examples are presented:

這首演奏曲表現出和諧的氣氛. (E2)
The piano melody performs harmonious mood.

Based on the matching algorithm, two substring CH_L and CH_R of polyphone 中 with 5 character length for E2;

CH_L ="這首演奏曲表現出和",

CH_R=" 和諧的氣氛".

Method 2: Language Model (LMs)

The experiment of three models unigram、bigram、trigram are implemented. Bigram reaches 92.58%, which is highest among three models.

Method 3: Voting Scheme

1)Winner take all： Three models; unitoken, betoken and tritoken are generated. Bitoken achieves PR of 90.17%.

2)Preference： Three models; unitoken, betoken and tritoken are generated. Bitoken achieves highest PR of 92.72%。

3.3 Word 2007 precision rate

MS Office is the famous and well used package around world. MS Word 2007 translate on same testing sentences, the PR achieves 89.8% in average for same testing sentences.

3.4 Results Analysis

In the paper, preference, winner take all, and language Model are proposed. We compare these methods with MS Word 2007. Preference betoken achieves highest PR among these models and reaches 92.72%. It is apparent that our proposed methods are all superior to the Word 2007.

4. Two-pass Approach

In the section, we will describe the two-pass approach, in which there are two sequential phases are constructed to promote the performance for WSD. The methodology of two-pass processing can be found in related fields, such as Named Entity Identification [12] and machine translation [13] while little work for WSD issues. The two-pass unified approach has been proposed and integrated three principal methods to improve furthermore the performance.

4.1 Alternative Method

How can we improve furthermore the prediction rate of Chinese polyphones? While the categories with top two maximum probability or scores are much closer, it leads usually to justify wrongly the right target. Therefore we propose the unified approach to resolve such a situation.

As described above, the Preference of voting scheme method with highest PR (92.72%) will be the main method predicting the target while the bigram LMs with (92.58%) is the alternative.

4.2 Unified approach I

As shown in Figure 1, the Preference method with betoken is employed normally. While the difference between top two scores with respect to the polyphone is less than the threshold θ_1, the alternative method, betoken language model, will be activated. In such situation, It means that the confidence for predicting the categories based on the features used by Preference method should be lower. These two method are unified under the threshold θ_1. The precision rates outside testing are enhanced up to 93,32% under θ_1 with 0.05.

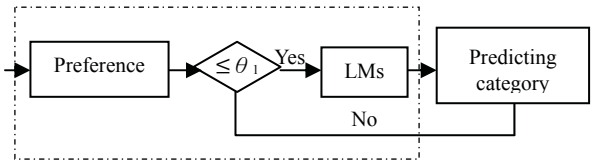

Figure 1: Unified approach I

4.2 Unified approach II

Although the bigram LMs is employed as the alternative shown in Figure 1, the situation described above will occur again. The winner-take-all with higher PR of voting scheme will be other alternative to resolve such situation, as shown in Figure 2. Same algorithm is used to the alternative under threshold θ_2. The precision rates are enhanced up to 94.3 under θ_2 with 0. It proved that the two-stage unified approach is effective and net PR reached totally 1.58% higher than the best result 92.72% in Section 3. It proves that the proposed approach can disambiguate the issue of Chinese polyphones and resolve effectively the text-to-phoneme conversion in on Chinese TTS system.

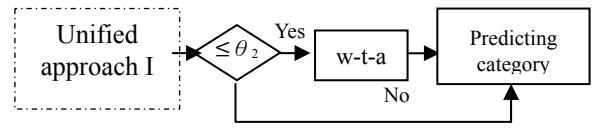

Figure 2: Unified approach II

5. Conclusion

The paper addresses issues of Chinese polyphones and disambiguity approaches for it. We proposed the two voting schemes, preference and winner take all scoring, for resolving the issues. The approach of unifying several methods in the paper will be discussed to enhance better performance. The predicted result is better than that of MS WORD 2007.

Furthermore the two-pass unified approach has been proposed and integrated three methods to improve the performance. The final precision of experimental result achieves 94.3%. It proves that the proposed approach can resolve effectively text-to-phoneme issue in Chinese TTS system.

In future, several researches topics are as follows:

- More lexical features, such as location and semantic information, used to enhance the precision rate (PRs).
- Improving the prosody of speech output.
- Improving the adaption of learning.

Reference

[1] Ruth Clark, 2002, Six Principles of Effective e-Learning: What Works and Why, Learning Solutions.

[2] Chang, Yi-Hsing Lu, Tsung-Yi, 2006, An Effective E-Learning System Based on Knowledge Management and Intelligent Agents, 2006 ICS, Taiwan.

[3] Ying-Hong Wang, Chu-Chi Huang, and Wen-Nan Wang, A Semantic-Aware Methodology Adapt to e-Learning Environment, Journal of Computers, Vol.17, No.3, October 2006, pp. 41~54

[4] Yanyan Li and Ronghuai Huang, 2006, Knowledge Science & Engineering Institute, Lecture Notes in Computer Science, Volume 4018.

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

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

[7] Ying-Hong Wang, Chu-Chi Huang, and Wen-Nan Wang, A Semantic-Aware Methodology Adapt to e-Learning Environment, Journal of Computers, Vol.17, No.3, October 2006, pp. 41-54.

[8] Dictionary of Chinese Idioms, MOE, R. O. C. website: http://140.111.34.46/chengyu/

[9]Delia Rusu, Blaž Fortuna, Dunja Mladenić, Sep. 2009, Improved Semantic Graphs with Word Sense Disambiguation, The Fifth International Conference on Knowledge Capture, California, USA,

[10]Paramveer S. Dhillon and Lyle H. Ungar, August 2009, Transfer Learning, Feature Selection and Word Sense Disambguation, ACL-IJCNLP 2009, Singapore. pp. 257–260.

[11]M. Barathi, S.Valli, 2010, Ontology Based Query Expansion Using Word Sense Disambiguation, International Journal of Computer Science and Information Security, Vol. 7, No. 2, February 2010, pp. 22-27.

[12] Jian Sun, Jianfeng Gao_, Lei Zhang, Ming Zhou, Changning Huang, 2002, Chinese Named Entity Identification Using Class-based Language Model, COLING-2002-044. Two pass

[13] Dekai WU, Pascale FUNGk, June 2009, Semantic Roles for SMT: A Hybrid Two-Pass Model, NAACL HLT 2009, pp. 13–16.