TEXT NORMALIZATION IN MANDARIN TEXT-TO-SPEECH SYSTEM

Yuxiang Jia[1,2], Dezhi Huang[2], Wu Liu[2], Yuan Dong[2,3], Shiwen Yu[1], Haila Wang[2]

[1]Institute of Computational Linguistics, Peking University, Beijing, China
[2]Speech and Natural Language Processing Unit, France Telecom R&D Beijing, China
[3]Beijing University of Posts and Telecommunications, Beijing, China

ABSTRACT

Text normalization is an important component in Text-to-Speech system and the difficulty in text normalization is to disambiguate the Non-Standard Words (NSWs). This paper develops a taxonomy of NSWs on the basis of a large scale Chinese corpus, and proposes a two-stage NSWs disambiguation strategy, Finite State Automata (FSA) for initial classification and Maximum Entropy (ME) classifiers for subclass disambiguation. Based on the above NSWs taxonomy, the two-stage approach achieves an F-score of 98.53% in open test, 5.23% higher than that of FSA based approach. Experiments show that the NSWs taxonomy ensures FSA a high baseline performance and ME classifiers make considerable improvement, and the two-stage approach adapts well to new domains.

Index Terms— Text-to-Speech (TTS), Text Normalization, Finite State Automata, Maximum Entropy Classifier

1. INTRODUCTION

Text normalization is a crucial component of text analysis in TTS systems. Real text contains many Non-Standard Words (NSWs), in that their properties can not be found in a dictionary, nor can their pronunciations be found by an application of "letter-to-sound" rules [1]. NSWs need to be normalized into their corresponding standard words and such a process is called text normalization. In English, number expressions, abbreviations, and acronyms are NSWs. Even sentence segmentation is a task of text normalization. For Chinese, non-Chinese words like numbers, symbols and alphabets need to be normalized into Chinese forms. A Non-Standard Word (NSW) could be converted into different standard words depending on both the local context and the text genre. So it is in general a very hard homograph disambiguation task [2]. In Nuance Vocalizer, over 20% of the core application code (line of code metric) is devoted to text normalization, and new input forms continue to be added [3].

Typical methods for text normalization are based on hand-crafted rules. But such hand-crafted rules are difficult to write, maintain and adapt to new domains. On the other hand, in view of homograph disambiguation, many machine learning methods are employed and have shown their advantages. Decision tree and decision list are used in English and Hindi text normalization [4]. Support Vector Machine is applied to Persian NSWs classification [5]. Winnow is used for homograph disambiguation in Thai text analysis [6].

However, most Chinese text normalization modules are rule based and preceded by word segmentation process, for Chinese text lacks white space between words [7][8]. Literature [9] adopts an external rule based Chinese text normalization method. It maintains over 400 external rules and makes use of word and part-of-speech information. Still others put word segmentation, named entity recognition and NSWs process into a unified framework [10][11].

The text normalization approach proposed in this paper does not need word segmentation process. Finite state automata detect NSWs from the real text and make an initial classification and then maximum entropy classifiers are used for further classification.

The rest of this paper is organized as follows. Section 2 describes the proposed approach in detail. Section 3 gives experiment results and analysis. Conclusions are given in section 4.

2. DESCRIPTION OF THE APPROACH

NSWs taxonomy is developed after a systematic investigation of a large scale corpus, the People's Daily Corpus. Based on this taxonomy, a three-layer normalization process is designed. Finite State Automata are used for NSWs detection and initial classification. Maximum entropy classifiers are applied for NSWs further classification. At last, Finite State Transducers are used for generation of standard words.

2.1. Taxonomy

The taxonomy of NSWs is the basis of text normalization. It defines categories of NSWs, according to which, NSWs are detected, classified and transformed. In Chinese real text, Arabic digits and some symbols are the major objects to be normalized. The taxonomy is based on one year of the People's Daily Corpus, which contains 300,277 digit strings.

Table 1. NSWs taxonomy based on input formats

Numbers	digits	110, …
	dot	1.29, 2000.9.10, 162.105.81.14, …
	hyphen	1998-2002, 2000-9-10, 4-3-2-1, …
	slash	1/3, 2000/9/10, …
	colon	10:15, 10:15:20, …
	suffixes	%,*万(ten thousand)*, qualifiers, ...
	range	100-200 *人(100 to 200 men)*, …
	others	'99, …
Symbols	-, /, :, ., ×, >, =, <, …	
Others	URL, Email, Alphabets, …	

Table 1 is a brief summary of the NSWs taxonomy. NSWs are firstly categorized by their formats. 95% of the 276,525 NSWs in the corpus are number expressions, including digit strings, various combinations of digit strings and symbols (dot, hyphen,

slash, colon, etc.), and digit strings with suffixes like Chinese qualifiers. Symbols are another category that needs to be converted and some symbols have many pronunciations. Normalizations of URL and Email addresses are determinate. English alphabet strings have their corresponding Chinese translations. All other rare NSWs will also be added to "Others" category. In total, 48 types of NSWs of different formats are included in the taxonomy. Of these types, some have determinate pronunciations, while some others not.

NSWs whose pronunciations are determined by formats are named as Basic NSWs (BNSWs), while those with ambiguities are called Ambiguous NSWs (ANSWs). Table 2 and table 3 give some examples of BNSWs and ANSWs respectively.

Table 2 shows the distributions of BNSWs in the People's Daily Corpus. As can be seen, BNSWs account for about 84% of all NSWs occurrences and "quantity" alone account for 55% of all those occurrences. That means, 84% of NSWs are pronunciation-determined by their formats and only 16% are ambiguous. In "quantity", suffixes like Chinese qualifiers and measures, are important signals to determine NSWs pronunciations.

Table 2. Examples of BNSWs

NSW class	Example	Percent
quantity	35 人*(35 people)*	55%
integer-unit	100 万*(one million)*	8%
percent	10%, 12.5%	6%
date	2007 年 10 月*(October 2007)*	4%
decimal-unit	1.5 万*(15 thousand)*	3%
basic-range	10-15 厘米*(10 to 15 cm)*	2%
years	50 年历史*(50 years history)*	2%
others	'99, Win32	4%

Table 3. Examples of ANSWs

NSW class	Say-as	Example
digits	digit by digit	2 米 11 *(2.11 meters)*
	integer	110
	sound-change	110
	English	p2p
a-hyphen	year-year	1998-1999
	telephone	010-12345678
	digit-digit	波音 737-200 *(Boeing737-200)*
	integer-integer	200-300
	rate	比分 2-3 *(Score is 2 to 3)*
	subtract	100-1=99
slash	fraction	1/3
	not pronounced	T65/66
	date	2001/01
colon	time	上午 10:15 *(10:15 AM)*
	rate	比分 10:15 *(Score is 10 to 15)*
year	some year	1999 年 *(the year 1999)*
	duration	1000 年 *(1000 years)*

Table 3 shows some categories of ANSWs and their possible ways of pronunciation. As can be seen, some NSWs have a high degree of ambiguities and their disambiguation needs both internal and contextual information.

Based on the above taxonomy, text normalization process is composed of three stages. The first stage uses Finite State Automata to detect NSWs from real text and make initial classification. The BNSW classification is finished in this stage. For an ANSW, the output of initial classification is used for subclass disambiguation. Maximum entropy classifiers are used in the subclass disambiguation module. When a NSW is labeled with a class tag, a Finite State Transducer transforms it into standard words. The process flow is outlined in Fig.1.

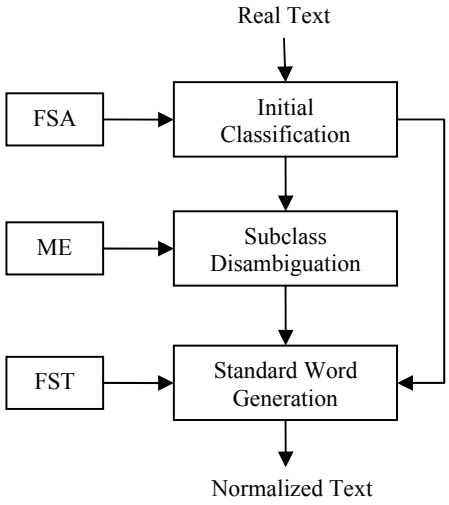

Fig.1. The flow chart of text normalization

2.2. NSW Detection and Initial Classification

Finite State Automata (FSA) are designed to detect NSWs and give an initial classification based on NSWs formats. Longer unit contains more information and thus has less ambiguity. So Maximum Match Strategy is adopted when nesting NSWs exist. It means the longest NSW is considered as a NSW, not any of its substrings.

FSA work on real text without word segmentation. When some Chinese suffixes are used in FSA, few segmentation problems may appear. For example, in NSW type "quantity", qualifiers, measures and some signal words are used as number suffixes, which are built into a Number Suffix List (NSL) like {人/天/时/…} *({People/Day/Hour/...})*. But in sentence "北京 1999 人才工程" *(Beijing 1999 talents project)*, FSA will detect "1999 人" *(1999 people)* as a NSW of number-suffix collocation. In fact, "1999" here does not refer to 1999 people, but the year 1999.

In order to solve this problem, we build a Secondary Suffix List (SSL) composed of words beginning with number suffixes, such as {人才/天津/时代/…} *({Talents/Tianjin city/Times/...})*. Thus, "1999 人才" *(1999 talents)*, instead of "1999 人" *(1999 people)* is recognized and further stage will give "1999" a class tag. NSL and SSL are extracted from an existing lexicon and the real text corpus.

2.3. Subclass Disambiguation

As is shown in table 3, ANSWs have different ways of pronunciation in different contexts. A subclass disambiguation process is needed to determine the true pronunciations in certain contexts. A maximum entropy classifier is built for each ANSW class.

Maximum Entropy Classifier

The maximum entropy framework agrees with everything that is known, but carefully avoids anything that is unknown. In other words, it estimates probabilities based on the principle of making as few assumptions as possible, under the constraints imposed. The probability distribution that satisfies the above property is the one with the highest entropy. It is of the following exponential form

$$P(y \mid x) = \frac{1}{Z(x)} \exp\{\sum_i \lambda_i f_i(x, y)\} \quad (1)$$

Where x is a history or context, y is the outcome or category, and $Z(x)$ is a normalization function.

$$Z(x) = \sum_y \lambda_i f_i(x, y) \quad (2)$$

The features used in the maximum entropy framework are binary. Any useful evidence sources can be incorporated into the model as features, without conditional independence assumption.

Here is an example of a feature function or indication function, which implies the fact that digit string "110" should be read digit by digit.

$$f_i(x, y) = \begin{cases} 1 & if \quad y = dd, nsw = 110 \\ 0 & otherwise \end{cases} \quad (3)$$

The training of maximum entropy model is to learn parameters λ_i . Parameter estimation methods include Generalized Iterative Scaling (GIS), Improved Iterative Scaling (IIS), L-BFGS and BLMVM, etc. Data smoothing methods include Gaussian prior, exponential prior, and inequality smoothing algorithm [12], etc. The fast parameter training method BLMVM and inequality smoothing algorithm are employed in our maximum entropy classifiers.

The Feature Templates

The set of classifiers have both public features and private features. Public features are shared by all classifiers while private features are designed specially for each classifier.

Public Features, as follows, are n-gram Character features within window size (equals 4 here),

Uni-gram: C_n *(n=-4,-3,-2,-1,0,1,2,3,4)*
Bi-gram: C_nC_{n+1} *(n=-4,-3,-2,-1,0,1,2,3)*
Tri-gram: $C_nC_{n+1}C_{n+2}$ *(n=-4,-3,-2,-1,0,1,2)*
4-gram: $C_nC_{n+1}C_{n+2}C_{n+3}$ *(n=-4,-3,-2,-1,0,1)*

Here, C_n can be a Chinese character, a digit string, an alphabet string, or a symbol.

$$C_n = \begin{cases} "num" & if \quad n \neq 0 \wedge C_n \_ is \_ digits \\ C_n & otherwise \end{cases} \quad (4)$$

If C_n in context is a digit string, it is substituted by "num", for a specific number may be not as indicative as the fact that it is a number.

Private features include some heuristic information for specific classifiers. Take the classifier for NSW type "digits" as an example. Its private features are as follows,

The number of digits in NSW
If it begins with zero
If it's preceded with alphabets
If it's followed with alphabets

While for NSW type "year", private features are about *the range of the number before "年(year)".*

2.4. Standard Word Generation

Standard word generation is the last module of text normalization. It is a generation step while former steps are analysis steps. The input of this module is NSW itself and its class tag. The output is its corresponding Chinese words. The conversion is a one-one correspondence and finite state transducers are applicable here.

3. EXPERIMENTS

Experiments are designed to test the performance of NSWs detection and classification. For standard word generation is a determinate transformation, experiments here reflect the performance of the whole text normalization process.

3.1. Corpora

In the whole process, only the 6 maximum entropy classifiers for ANSWs need to be trained. For each classifier, the training set is composed of all occurrences of that type of ANSW in the People's Daily Corpus. The training set sizes are ranging from 400 sentences to 4000 sentences. The closed test corpus is randomly extracted from the People's Daily Corpus. It contains 6986 sentences and 13468 NSWs. Each sentence on average has 1.93 NSWs. The open test corpus is collected from the internet. Web pages are of various domains like sports, digital products, military and Bulletin Board System, etc. It contains 6007 sentences and 12219 NSWs. Each sentence on average has 2.03 NSWs. The percentage of BNSWs in the closed test corpus is 83.59%, while in the open test corpus it is 74.77% (See table 4).

Table 4. Distribution of NSWs in test corpora

	Sentences	NSWs	BNSWs	ANSWs
Closed	6986	13468	83.59%	16.41%
Open	6007	12219	74.77%	25.23%

In the corpora, a pair of brackets "[" and "]" are used to indicate a NSW. Class tag follows "]" and if needed, a subclass tag follows class tag with a "/" as a separator. For example, "拨打110" *(Dial 110)* is tagged as "拨打[110]digits/dd", where "110" is a NSW of class "digits", and it should be read digit by digit. The annotation is conducted by two annotators, a graduate student majoring in Chinese language and the first author. The inter-annotator agreement is measured by the Kappa statistic (K). We randomly extract 250 samples of the most problematic NSW type "digits", and get K=0.9830, which is a very good agreement.

3.2. Experiments and Results

Only if a NSW is correctly recognized and classified, it is counted as a correctly tagged NSW. Evaluation criteria are Precision (*P*), Recall (*R*) and F-score (*F*). Where,

$$P = \frac{\#correctly\_tagged\_NSW}{\#auto\_tagged\_NSW} \quad (5)$$

$$R = \frac{\#correctly\_tagged\_NSW}{\#real\_NSW} \quad (6)$$

$$F = \frac{2 \times P \times R}{P + R} \quad (7)$$

As the baseline, we use FSA for initial classification and simply label ANSWs with their major subclass tags. In table 5, baseline (FSA) achieves F-scores of 96.99% in the closed test and 93.30% in the open test. BNSWs, which are always the majority of NSWs, ensure a good baseline performance across domains.

Table 5. Overall performance

	Closed (%)			Open (%)		
	P	*R*	*F*	*P*	*R*	*F*
FSA	96.99	96.99	96.99	93.30	93.30	93.30
FSA+ME	99.96	99.96	99.96	98.53	98.53	98.53

When Maximum Entropy (ME) classifiers are used for ANSWs, F-score gets improved by 2.79% in closed test and 5.23% in open test. It shows that ME is effective for NSWs disambiguation and the text normalization module in this paper adapts well to new domains.

In table 5, precision equals to recall, which means that the number of auto-tagged NSWs equals to the number of real NSWs. Experiments prove that FSA detect all NSWs correctly. Errors are only introduced in the subclass disambiguation stage. The micro average precisions of subclass disambiguation in the closed test and open test are 99.73% and 94.23% respectively.

3.3. Error Analysis

Some errors occur in experiments. For example, "邮政编码 [250100]digits/dd， 应读作 [25]digits/in 、 [01]digits/dd 、 [00]digits/dd" *(Post code 250100 should be read as 25,01,00)* is a number sequence error. "25" should be tagged as [25]digits/dd, which means "25" is to be read digit by digit. The tag of "25" here depends on its neighboring tags. To utilize neighboring tags, a sequence labeling model may be applicable, such as Maximum Entropy Markov Model (MEMM) or Conditional Random Field (CRF) model.

4. CONCLUSION AND FUTURE WORK

This paper makes an extensive investigation of Chinese text normalization. NSWs taxonomy is developed based on a large scale corpus. After a systematic analysis of the taxonomy, a two-stage NSWs classification strategy is proposed, finite state automata for initial classification and maximum entropy classifiers for further classification. Experiment results show that this approach achieves a good performance and generalizes well to new domains. In addition, this approach is character-based, no need of word segmentation preprocess.

To solve number sequence errors, some heuristic rules or sequence labeling models, like MEMM or CRF will be considered. More knowledge sources will be used to classify some symbols. For example, a location name list is useful to classify "-" in "北京-上海" *(Beijing-Shanghai)* while a surname list will help to determine pronunciation of "×" in "陈×" *(Person with surname Chen)*. Weakly supervised or unsupervised learning methods will be studied to reduce human involvement.

ACKNOWLEDGEMENTS

The work in this paper is supported by Funds for Key Project of Chinese National Programs for Fundamental Research and Development (No.2004CB318102). We are grateful to the anonymous reviewers for their helpful advice to improve the paper.

REFERENCES

[1] Richard Sproat, Alan Black, Stanley Chen, Shankar Kumar, Marsi Ostendorf, and Christopher Richards, "Normalization of Non-Standard Words," *Computer Speech and Language*, 15(3):pp. 287-333, 2001.

[2] David Yarowsky, "Homograph Disambiguation in Text-to-Speech Synthesis," In Jan van Santen, Richard Sproat, Joseph Olive, and Julia Hirschberg (eds.), *Progress in Speech Synthesis*, Springer, New York, pp.159-175, 1996.

[3] Andrew Breen, Barry Eggleton, Peter Dion, and Steve Minnis, "Refocusing on the Text Normalization Process in Text-to-Speech Systems," *In Proc. ICSLP 2002*, pp. 153-156, 2002.

[4] K. Panchapagesan, Partha Pratim Talukdar, N. Sridhar Krishna, Kalika Bali, and A.G. Ramakrishnan, "Hindi Text Normalization," *In Proc. KBCS 2004*, pp.19-22, 2004.

[5] M.H.Moattar, M.M.Homayounpour, and D.Zabihzadeh, "Persian Text Normalization Using Classification Tree and Support Vector Machine," *In Proc. ICTTA 2006*, pp.1308- 1311, 2006.

[6] Virongrong Tesprasit, Paisarn Charoenpornsawat and Virach Sortlertlamvanich, "A Context-Sensitive Homograph Disambiguation in Thai Text-to-Speech Synthesis," *In Proc. HLT-NAACL 2003*, pp.103-105, 2003.

[7] Chilin Shih, and Richard Sproat, "Issues in Text-to-Speech Conversion for Mandarin," *Computational Linguistics and Chinese Language Processing*, 1(1): pp.37-86, 1996.

[8] Min Chu, Peng Hu, Yong Zhao, Zhengyu Niu, and Eric Chang, "Microsoft Mulan--a bilingual TTS system," *In Proc. ICASSP 2003*, pp.264-267, 2003.

[9] Zhigang CHEN, Guoping HU, and Xifa WANG, "Text Normalization in Chinese Text-to-Speech System," *Journal of Chinese Information Processing*, 17(4): pp.45-51, 2003.

[10] Jianfeng Gao, Mu Li, Changning Huang, and Andi Wu, "Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach," *Computational Linguistics*, 31(4): pp.531-574, 2005.

[11] Guohong Fu, Min Zhang, Guodong Zhou, and Kang-Kwong Luke, "A Unified Framework for Text Analysis in Chinese TTS," *In Proc. ISCSLP 2006*, pp.200-210, 2006.

[12] Jun'ichi Kazama, and Jun'ichi Tsujii, "Evaluation and Extension of Maximum Entropy Models with Inequality Constraints," In Proc. EMNLP 2003, pp.137-144, 2003.