Spontaneous Mandarin Speech Understanding Using Utterance Classification: A Case Study

Yun-Cheng JU, Jasha DROPPO
Speech Research Group
Microsoft Research
Redmond, WA. USA
{yuncj, jdroppo}@microsoft.com

Abstract—**As speech recognition matures and becomes more practical in commercial English applications, localization has quickly become the bottleneck for having more speech features. Not only are some technologies highly language dependent, there are simply not enough speech experts in the large number of target languages to develop the data modules and investigate potential performance related issues. This paper shows how data driven methods like Utterance Classification (UC) successfully address these major issues. Our experiments demonstrate that UC performs as well as or better than hand crafted Context Free Grammars (CFGs) for spontaneous Mandarin speech understanding, even when applied without linguistic knowledge. We also discuss two pragmatic modifications of the UC algorithm adopted to handle multiple choice answers and to be more robust to feature selections.**

Keywords-CFG; Utterance Classification; Mandarin Language Understanding; Spoken Language Understanding (SLU)

I. INTRODUCTION

American English speech recognition techniques are mature and practical enough to enable a variety applications and devices. Applications range from voice dialing and voice search on phones, call routing [1] and voice menus on telephony services, and voice commands on smart phones, PCs, in-car infotainment systems [2], or even game consoles [3]. However, even in American English, we see a tremendous challenge for the industry because of the unprecedented deployment of speech technologies in the hands of non-speech experts.

The recent globalization trend requires many localized versions when a product or service ships, and the speech recognition component is arguably the single most difficult and expensive feature to be localized. Even when the engine providers have the acoustic models for the target language, it's still too difficult for application developers to develop, diagnose, and optimize the task dependent language models and other spoken language understanding components. Most developers don't have enough understanding in both the technology and the target language to get satisfactory performance.

As a result, localization becomes the bottleneck to providing speech features. In fact, there have been several occurrences within the company where promising speech features were called off because of the business groups were discouraged after realizing the challenges of localization.

Localization to Mandarin poses many specific challenges to speech recognition and spoken language understanding. Not only because it is a tonal language and the word units are less well defined, it also has distinctly different syntactic structure from English.

This paper seeks language independent technologies and recipes for spontaneous Mandarin speech understanding. The authors argue the need for these methods even if they slightly underperform the traditional approaches which require language/speech expertise. A similar approach can be applied to many languages and lower the barriers to having more speech recognition features.

The rest of this paper is organized as follows: Section 2 describes the common practice of developing and tuning Context Free Grammars (CFGs) and highlights the localization challenges; Section 3 explains the Utterance Classification (UC) approach and demonstrates a few pragmatic solutions which do not require the linguistic and technical knowledge; Section 4 provides evaluation of the proposed approaches, and Section 5 concludes the work and points to future work.

II. CONTEXT FREE GRAMMARS (CFGS)

A. *Common Practice of Developing and Localizing CFGs*

The traditional approach, which is still widely used in the industry, is to write a CFG to capture the expected user utterances [4]. While this approach may look easy, it does not scale well [5] and is difficult to localize for several reasons.

The utterances are factorized, according to the syntactic and semantic structures, into phrases and word chunks before they can be grouped into the CFG rules. A common localization mistake is to translate the word chunks separately into the target language. The CFG translated this way usually does not perform well and is more likely to be syntactically incorrect.

Data collection in the target language is required to collect the whole utterances, preferably from the real usage, to reflect the correct distribution. While this process is time consuming and costly, we argue it is the easiest step because it doesn't require linguistic and technical expertise.

Due to linguistic variations, there are many ways people may say things. In addition, the size of data collection is usually small because of its high cost. As a common practice, developers usually manually generalize the observed utterances into artificial rules hoping the increased coverage causes higher performance.

We argue this "manual generalization" practice is most dangerous and the obstacle of the entire localization process, especially because it requires the linguistic knowledge in the target language.

B. Two Spontaneous Mandarin Understanding Examples

To illustrate the practice and evaluate the performance, we use two simple spontaneous Mandarin speech understanding tasks from the MAT2500 speech corpus [6] where each subject was asked to tell his/her language background and education level.

The first task is to identify the subject's education level from the given prompt "您的教育程度是小學,國中,高中,專科,大學,或研究所程度?" (*What's your education level? Elementary School, Middle School, High School, College, University, or the level of Graduate School?*) Even though it is a simple multiple choice question, there are still many varieties in the answers after excluding the common synonyms for each level (e.g., "博士", "高職", "國小", "五專", etc). Some popular answers and the corresponding English translations are listed in Table 1.

While it is easy to tag each answer with the correct label, it is not a simple decision to find the best generalization from the examples, especially for developers not fluent in Mandarin. The small size of the samples also makes it more difficult to separate outliers from valuable new paterns. Worse, researchers [7] found that adding more user response alternatives sometimes actually causes severe performance degradation due to the increased confusion in the recognition. A simplified version of the handcrafted CFG configuration is illustrated in Figure 1 without showing the weights.

Table 1: Less Expected Answers Observed from the Education level Task

Less Expected Answers	English Translation
研究所程度	Level of Graduate School
大學畢業	Graduated from University
大專	University or College
大學二年級	University Sophomores
國中肄業	Dropped out from Middle School
目前是高中	Currently in High School
現在念國中	Currently in Middle School

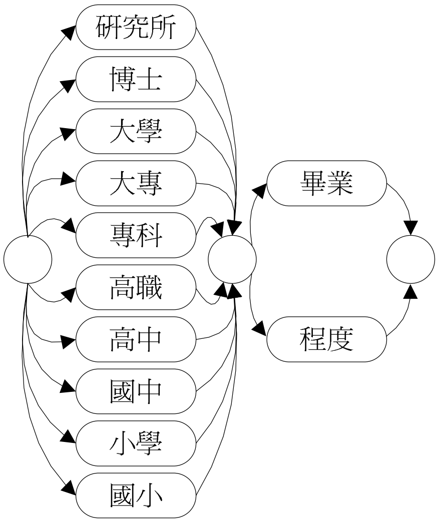

Figure 1: Handcrafted CFG for Education Level Task

The second task is to identify the language/dialect the subject and his/her mother speaks. In order to have more data samples, we merged the first two questions, namely "您日常講的是國語,閩南語,客家語或是其他那種語言?" and "您母親講的是國語,閩南語,客家語或是其他那種語言? (*What's the language or dialect you speak every day?* and *What's the language or dialect your Mom speaks? Mandarin, Taiwanese Hokkian, Kakka, or other languages?*) Certainly, the answer set is more complex as there are many bi-lingual and even tri-lingual speakers.

The need to accommodate potential multiple answers from the list of choices significantly increases the complexity and difficulty of CFG authoring. There are simply too many ways people answer this question even after merging the common synonyms for each dialect (e.g., "台語", "客家話", "閩(mi3)南語") and the many varieties of the conjunctions for "and" and "or" (e.g., "和(han4)", "跟", "及", "以及", "或", "或是", "和(her2)", "加", "與", "還有", etc). There is no difference between "both" and "all" in Mandarin (i.e., "都有") and the conjunction word can also be skipped.

The less expected yet popular answers observed from the corpus is listed in Table 2. Even for speech experts fluent in Mandarin, it's not obvious what the best generalization rule is. A simplified version of the handcrafted CFG configuration for this task is illustrated in Figure 2. Clearly, more data driven and language independent approach is needed to overcome the localization bottleneck.

Table 2 : Popular Answer Examples for the Dialect Task

Popular Answers	English Translation
國語閩南語	Mandarin, Taiwanese
國語和閩南語	Mandarin and Taiwanese
國語閩南語客家語	Mandarin, Taiwanese, Kakka
國語閩南語都有	Both Mandarin and Taiwanese

都有	All of them
三種都有	All of the three
國台語並用	Mandarin and Taiwanese together
大部份是閩南語	Mostly Taiwanese
有的時候講國語有的時候講客家話	Sometimes speak Mandarin Other times speak Kakka
國語閩南語交替使用	Switch between Mandarin and Taiwanese
以國語為主閩南語為輔	Mainly Mandarin but also Taiwanese
台灣國語	Mandarin with Taiwanese accent
三種語言都使用	All of the three
廣東話	Cantonese
嗯閩南語	(filler) Taiwanese
國語啊	Mandarin (filler)

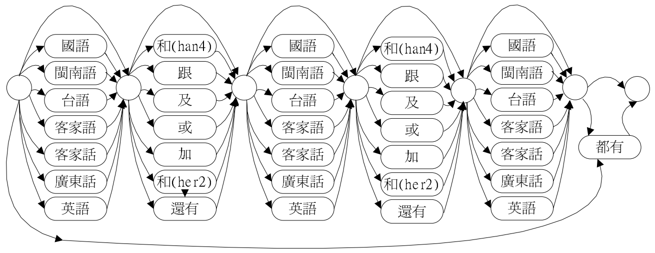

Figure 2: Handcrafted CFG for Dialect Task

III. Utterance Classification

Utterance Classification has been widely applied to the task of call routing. It is statistical and data-driven in nature and is typically used to classify natural spoken responses to open-ended prompts like "How may I direct your call?"

The standard UC process is depicted in Figure 3. The speech recognizer uses a language model, most likely a statistical n-gram, to transform the speech utterance into words, or other lexical units [8]. It is followed by a text classifier to further categorize the recognized text into a fixed set of semantic concepts.

This procedure requires the training of both the language and classifier models. We argue the use of two separate modules can automatically generalize from the training sentences without the manual process and, in turn, eliminate the need for linguistic and speech expertise. In addition, our previous study [9] shows we can train an SLM to get reasonable Semantic Classification Errors (SER) from a very small training set. Unsupervised language model adaptation can also be used to use un-transcribed audio data [10].

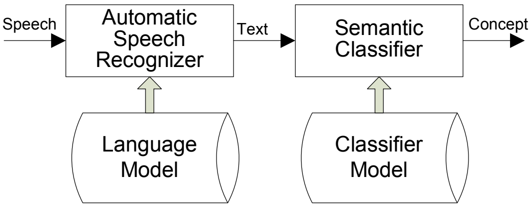

Figure 3: Utterance Classification (UC) System Diagram

A. *Statistic Language Model*

In general, a better language model helps the speech recognizer to reduce the Word Error Rate (WER) in transcribing the speech into text. A lower WER means cleaner transcriptions for the semantic classifier, which usually improves the Classification Error Rate (CER). However, research has shown the configuration that achieves the lowest WER might not always gives the lowest CER [11] and suggests researchers focus on building language models which produce more consistent transcriptions, but not necessarily lower WER.

The definition of "word" doesn't exist naturally in Mandarin and some other languages including Japanese as well. Even though artificial word units are used in Mandarin dictation systems to improve WER [12], they are statistically induced from the training corpus and might not be practical to ordinary speech understanding tasks.

In this paper, we examine the use of Chinese character, or syllable as the SLM word unit. Not only because it's simpler and more natural, but also it has some benefits of better generalization in the next classification phase as suggested from our previous work [13]. In fact, homonyms are also merged and we use Pin-Yin as our final word unit. As an example, the recognized text for the utterance "國台語並用" is "guo2 tai2 yu3 bing4 iong4" (5 word units).

This process also significantly simplifies the pronunciation lexicon.

B. *Classification Model*

A classification model serves as a routing table to map a text string into a fixed set of concepts. Both Maximum Entropy [14] and Vector Space Model [15] based classifiers have been widely used in call routing applications. The best choice of classifiers is beyond the scope of this project. In our research, we use the exact classifier setup from our previous work [13] using bigram syllable features. The routing matrix is initialized using the TF-IDF formula [16] and then discriminatively trained based on Minimum Classification Error (MCE) criterion using procedures similar to [1] to guarantee a minimized CER.

The intuition of using character/syllable but not words in our approach is to achieve a more robust algorithm against poorly chosen word units and generalizes better from limited data. For example, only seeing "客家語" in the training sentences might be sufficient to understand the variation of "客家話", or just "客家".

C. *Multiple Answers from the List of Choices*

While the SLM addresses the recognition challenges the CFG approach is facing, a monolithic classifier is inappropriate for prompts which require recognition of multiple answers from a list, such as “Which pizza toppings do you want?” There are two main issues.

First, the choices are mostly independent. Standard feature selection might not provide sufficient information and some long distance features are required, which is a situation we definitely want to avoid. To illustrate the significance of this problem even when the “best” possible word units are used, a few examples (in English translation) in Table 3 illustrate the bigram features and the difficulties they face in a monolithic configuration.

Table 3: Bigram Features in the Dialect Task in English

Example	Tag	Bigram Features
Mandarin and Taiwanese	MT	“<s> Mandarin”, “Mandarin and”, “and Taiwanese”, “Taiwanese </s>”
Taiwanese or Mandarin	MT	“<s> Taiwanese”, “Taiwanese or”, “or Mandarin”, “Mandarin </s>”
Kakka or Mandarin	MK	“<s> Kakka”, “Kakka or”, “or Mandarin”, “Mandarin </s>”
Mandarin and Kakka	MK	“<s> Mandarin”, “Mandarin and”, “and Kakka”, “Kakka </s>”
Kakka or Taiwanese	TK	“<s> Kakka”, “Kakka or”, “or Taiwanese”, “Taiwanese </s>”
Taiwanese and Kakka	TK	“<s> Taiwanese”, “Taiwanese and”, “and Kakka”, “Kakka </s>”

In addition, as the size of the choices increases linearly, the combinations increase exponentially. It is likely that there might be no training examples for some combinations.

We propose the use of a federation of several independent binary classifications for each choice separately, and then combine the results together. This pragmatic solution generalizes the training sentences better and is less sensitive to the choice of word units. Table 4 shows the same examples sentences and the corresponding bigram character features. It is clearly illustrated that the new approach is more capable of identifying the task correctly.

Table 4: Character Bigram Features in the Dialect task

Example	Tag M	Tag T	Tag K	Bigram Features
國語和閩南語	+	+	-	“<s>國”, “國語”, “語和”, “和閩”, “閩南”, “南語”, “語</s>”
閩南語或國語	+	+	-	“<s>閩”, “閩南”, “南語”, “語或” , “或國”, “國語”, “語</s>”
客家話或是國語	+	-	+	“<s>客”, “客家”, “家話”, “話或”, “或是”, “是國”, “國語”, “語</s>”
國語客家語	+	-	+	“<s>國”, “國語”, “語客”, “客家”, “家語”, “語</s>”
客家話或台語	-	+	+	“<s>客”, “客家”, “家話”, “話或”, “或台”, “台語”, “語</s>”
台灣話和客家話	-	+	+	“<s>台”, “台灣”, “灣話”, “話和”, “和客”, “客家”, “家話”, “話</s>”

D. *Whole Utterance Cache Model*

The step of feature selection, unfortunately, is still heuristic and depends on experiences in the UC technique. We observed a few situations where even some training sentences didn’t get classified correctly. In order to avoid this uncertainly and inconvenience, we adopted a whole utterance cache model to override the decision from the classifier if the exact same utterance has been observed in the training set.

IV. Evaluation

We measure the semantic accuracy and conducted the test using the MAT2500 corpus. We randomly partition the 2232 speakers 50 – 50 into either the training set or test set and discarded all utterances labeled as “unusable”.

A. *Acoustic Model Training*

The acoustic model was trained on a set of 72,493 utterances from the corpus. A total of 410 context-independent, tone-independent syllable HMM models were trained. Each contained three states with eight Gaussian mixture components per state. Additionally, two silence models were included, as well as a garbage model called ENGLISH used to align with and ignore any non-Mandarin words in the training set. The final model contains 9909 mixture components in 1650 acoustic states.

B. *The Education Level Task*

There are 1091 utterances in the training set and 1073 utterances in the test set for the education level task. We were only interested in identifying the main categories of the education levels and ignored the additional information like the grade or whether the subject dropped out before finishing the school. We also created another category “大專” as it could be either “大學” or “專科”.

We tested the best CFG performance with 3 SLM/UC approaches: Syllable-based SLM/UC (S), Syllable-based SLM/UC plus whole utterance cache model (S+C), and Word-based SLM/UC plus whole utterance cache model (W+C).

Table 5 shows the results. The SLM/UC(S) approach made only 2 more mistakes compared with the CFG approach. Most of the mistakes are mis-recognitions, especially between “小學” and “大學”, and can’t be recovered. It also only made one more mistake compared with the more complicated word-based system. It is clear that the syllable-based SLM/UC approach is almost as good as using a custom-built CFG, but without the need for language expertise.

Table 5: Comparison of Accuracy for Education Task

Configuration	CFG	UC(S)	UC(S+C)	UC(W+C)
Accuracy (%)	98.32	98.13	98.13	98.23

C. *The Dialect/Language Task*

Because we use two questions together, there are 2059 utterances in the training set and 2048 utterances in the test set for the dialect task. Since there is usually one occurrence of the so many other dialects/languages (e.g., "山東話", "湖北話", "法文", etc) in the entire corpus, we decided not to try to recognize them separately but identify them as "其他" (Other, or "O").

We tested the best CFG performance with three SLM/UC approaches: Syllable-based SLM/UC (S), Syllable-based SLM/UC plus whole utterance cache model (S+C), and Word-based SLM/ "Word Spotting" plus whole utterance cache model (WP+C). For each SLM/UC approach, we also tested three configurations: either using a single classifier (monolithic), or the combination of four separate binary classifiers, one for each dialect (federation).

Table 6 shows the results. The syllable-based SLM/UC approach slightly under-performed the more sophisticated word-based system but was able to out-perform the best CFG configuration. Using a combination of four separate binary classifiers outperforms the monolithic approach and the whole utterance cache model recovered a few classification errors.

Table 6: Comparison of Accuracy for dialect Task

Accuracy %		CFG	UC(S)	UC(S+C)	UC(WP+C)
Monolithic		98.58	98.34	98.73	99.12
Federation	Mandarin		99.22	99.41	99.60
	Taiwanese		99.61	99.61	99.56
	Kakka		99.56	99.56	99.80
	Other		99.56	99.56	99.70
	overall		98.58	98.78	99.12

V. Conclusion

The unprecedented deployment of speech technologies in the hands of non-speech experts and the requirement of globalization poses a tremendous challenge for the industry. Simple, reliable, data driven and cost effective methods must be found to deliver speech recognition and understanding in multiple languages.

In this paper, we demonstrated that Utterance Classification with syllable based statistic language model met or exceeded the best handcrafted CFG for two spontaneous Mandarin understanding tasks without deep language specific linguistic knowledge. Pragmatic solutions were also proposed to overcome the exponential combination of multiple answers and to be more robust to classification errors.

The future work includes developing more language independent data driven methods and applying the syllable-based SLM/UC to more languages.

Acknowledgment

The authors would like to thank Prof. Jen-Tzung Chien and Prof. Jyh-Shing Roger Jang for suggesting the speech corpus and many useful discussions.

References

[1] Hong-Kwang Jeff. Kuo and Chin-Hui Lee, "Discriminative training in natural language call routing," in Proc. of ICSLP, 2000.

[2] http://www.fordvehicles.com/technology/sync

[3] http://en.wikipedia.org/wiki/Kinect

[4] Intervoice Training Document - Voice User Interface Design - Speechworks 7.0 OSS/OSR and Nuance 8.0 - Speech Forms, Intervoice Inc.,2004.

[5] D. Suendermann, K. Evanini, J. Liscombe, P. Hunter, K. Dayanidhi, R. Pieraccini, "From rule-based to statistical grammars: continous improvement of large-scale spoken dialog systems", IEEE International Conference on Acoustics, Speech and Signal Processing, 2009.

[6] Hsiao-Chuan Wang, "MAT -- A project to collect Mandarin speech Data through telephone networks in Taiwan," Computational Linguistics and Chinese Language Processing vol.2, no.1, February 1997.

[7] M. Balakrishna, D. Moldovan, E.K. Cave, "Automatic creation and tuning of context free grammars for interactive voice response systems", Proceedings of IEEE NLP-KE' 05.

[8] H. Alshawi, "Effective utterance classification with unsupervised phonotactic models." In the Proceedings of HLTNAACL, Edmonton, Canada, 2003.

[9] Yun-Cheng Ju, Ye-Yi Wang, and Alex Acero, "Call analysis with classification using speech and non-speech features", in the International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 2006.

[10] M. Bacchiani and B. Roark, "Unsupervised language model adaptation." In the Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, China, 2003.

[11] Y.-Y. Wang, J. Lee, and A. Acero, "Speech utterance classification model training without manual transcriptions," in Proc. Int. Conf. Acoustic., Speech, Signal Process., Toulouse, France, 2006.

[12] M. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin broadcast news speech recognition", In Proc. ICSLP, 2006.

[13] Xiaoqiang Xiao, Jasha Droppo, and Alex Acero, "Information retrieval methods for automatic speech recognition", in IEEE ICASSP, IEEE, 2010.

[14] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra, "A maximum entropy approach to natural language processing." Computational Linguistics, vol. 22, 1996.

[15] J. Chu-Carroll and B. Carpenter, "Vector-based natural language call routing," Computational Linguistics, vol. 25, no. 3, 1999.

[16] G. Solton and C. Buckley, "Term weighting approaches in automatic text retrieval", Information Processing and Management, 24(5):513-523, 1988.