A Word-Probabilistic Interface to Dialogue Modules

Alex Chengyu Fang, Weigang Li and Jonathan Webster
Department of Chinese, Translation and Linguistics
City University of Hong Kong
83 Tat Chee Avenue Kowloon, Hong Kong
{acfang,weiganli,ctjjw}@cityu.edu.hk

Abstract

A telephony dialogue system is described that performs speech-driven terminological translation. In particular, a novel approach is presented and discussed that is designed to probabilistically choose from a set of predefined, plan-based dialogue modules in order to maximise system usability. It is shown that words of different lengths, defined in terms of characters and syllables, demonstrate predictable degrees of recognition accuracy by the ASR engine. When expressed probabilistically, such varying degrees can be effectively used for the choice of appropriate dialogue modules. The novelty of this work is the measurement of word correct rate (WCR) as a function of grammar size and word length, expressed as WCR based on characters (*WCR-C*) and WCR based on syllables (*WCR-S*). The experimental results show that *WCR-C* and *WCR-S* can offer strong support in the development of an effective dialogue system, enhance dialogue flow and improve usability.

1 Introduction

Man-machine dialogue systems make use of different dialogue strategies to clarify user intent and to respond in an appropriate way. Typically, a dialogue system comprises different dialogue modules that handle different situations in the process of intension clarification. In speech-driven systems in practice, this boils down to the accuracy of the automatic speech recognition (ASR) system and how the system responds to different situations. For example, given the following dialogue turn:

System: Which term would you like to translate?
User: Gearbox.

the ASR engine will have a Boolean return. In the case of a positive one, the dialogue system will respond:

System: You said 'gearbox'. Its translation in Chinese is 齿轮箱.

With a negative ASR return, the system will say something like:

System: I'm sorry. Could you please repeat?
User: I said gearbox.

To enhance system usability, a third scenario is often necessary, where the caller is asked to confirm the ASR return:

System: Did you say gearbox?
User: Yes.

As can be seen, the three dialogue modules are components of an interactive session that attempts to verify the semantics of caller intent. A spoken dialogue system is typically configured to make use of the confidence level provided by the ASR engine in order to decide which module to opt for. There is also work to combine a second confidence score that represents an estimation of the mapping between the ASR result and user intention.

In this article, we report our work that aims to establish a third confidence score that is estimated externally on the linguistic string uttered by the speaker. Simply put, the score is an estimation of the probable ASR error rate according to the length of the word. The proposal of this additional confidence score is necessary since the other two scores do not take into account the fact that words of different lengths tent to have a different impact on the ASR engine. In addition, the size of ASR language models or grammars also has a significant impact on ASR performance. Our work to be reported here is therefore concerned with ASR evaluation according to two parameters: word length and grammar size.

Effective evaluation is an important task in spoken language dialogue systems (SLDS). Generally speaking, there are two purposes. One is to compare performance of different systems. The other is to improve the evaluated system itself. Different methodologies have been proposed to evaluate components in spoken language dialogue systems, such as Word Error Rate (WER) and weighted keyword error rate (WKER) (Nanjo and Kawahara, 2005; Hildebrandt et al., 1996). Higashinaka and colleagues describe a method for creating an evaluation measure for discourse understanding in spoken dialogue systems (Higashinaka et al., 2004). There is also a focus on user-related issues, such as user reactions to SLDS, user linguistic behaviour or major factors which determine overall user satisfaction (Walker et al., 1997; Walker et al., 2001; Hartikainen et al., 2004). There is increased focus on usability evaluation of SLDS in recent years (Dybkjr and Bernsen, 2001; Park et al., 2007) and metrics have been proposed, such as modality appropriateness, naturalness of user speech, and output voice quality.

All these methods are concerned with objective or subjective criteria of SLDS (Larsen, 2003). They aim to describe the system performance on the whole or part. Additionally, they all evaluate SLDS beyond the word level. This article discusses the fine evaluation of word-level performance in terms of word correct rate (WCR) and argues that there is much useful information at the word level that can improve SLDS performance effectively and efficiently.

2 Motivation

RAMCORP is a project that aims at the design and construction of a telephony dialogue system that provides on-the-spot machine translation of terminologies of a pre-defined domain. The interactive dialogue system uses Nuance, an off-the-shelf automatic speech recognition system, for the recognition of key words. In order to maximize transaction completion rate, RAMCORP will consist of several dialogue modules with different dialogue turns. A novelty of the project is to dynamically determine which dialogue to opt for according to the word being recognized. To achieve this, empirical experiments were carried out to ascertain the word correct rate (WCR) according to grammar size and word length. While it is common practice to measure WCR according to grammar size, the measurement of WCR as a function of word length has not been widely reported before. We define word length in two different ways: according to number of characters (WCR-C) and according to number of syllables (WCR-S). Results of the empirical experiments will ultimately inform the design of a formula that dynamically calculate the likelihood of a word being correctly recognized according to the three parameters, i.e., grammar size, number of characters, and number of syllables. Effectively, the system will be able to predict the likelihood of a word being correctly recognized and choose a corresponding dialogue module according to this likelihood.

This paper will focus on the empirical experiments that were carried out to establish the baseline statistics for Nuance. It will first of all report data selection including the selection of participating subjects and the selection of words that were used to form mock-up grammars of various sizes. It will then evaluate the ASR performance and report the resulting WCRs according in and discuss major findings.

3 Experiments and Analysis

3.1 Experimental setting

The off-the-shelf application used in this paper is Nuance Voice Platform (NVP). A demo dialogue system with word grammar rules is built for evaluation. Four grammars were constructed, consisting of only words to be recognized without any context cues. They respectively include 500, 1000, 2000 and 4000 words randomly selected the machine readable Collins English Dictionary. Twenty subjects as evaluators were invited to participate in the experiment. Each was asked to read four groups of 50 words randomly selected from the four grammars.

We thus obtained 20 sets of recognition results for grammars of four different sizes. The results of the experiment are summaries in Table 1.

S	WCR_{500}	WCR_{1000}	WCR_{2000}	WCR_{4000}	M
1	68.0	60.0	60.0	48.0	59.0
2	48.0	62.0	44.0	44.0	49.5
3	64.0	70.0	62.0	52.0	62.0
4	78.0	84.0	64.0	62.0	72.0
5	72.0	64.0	66.0	60.0	65 .5
6	62.0	60.0	46.0	44.0	53.0
7	84.0	58.0	58.0	50.0	62.5
8	88.0	66.0	76.0	64.0	73.5
9	72.0	80.0	56.0	50.0	64.5
10	68.0	52.0	58.0	58.0	59.0
11	64.0	64.0	56.0	50.0	58.5
12	74.0	60.0	50.0	40.0	56.0
13	58.0	58.0	64.0	54.0	58.5
14	72.0	44.0	66.0	44.0	56.5
15	82.0	74.0	76.0	50.0	70.5
16	82.0	78.0	82.0	58.0	75.0
17	76.0	72.0	74.0	56.0	69.5
18	82.0	84.0	62.0	58.0	71.5
19	78.0	68.0	68.0	68.0	70.5
20	76.0	70.0	76.0	58.0	70.0
M	72.4	66.4	63 .2	53.4	63.85

Table 1: Word correct accuracy and grammar size

3.2 Evaluation of WCR on Grammar Size

The most popular evaluation metric of ASR is Word Error Rate (WER), which is the minimum string edit distance between the correct transcription and the recognition hypothesis. There will be some new measures to propose to finely evaluate the dialogue system. In order to distinguish traditional WER, Word Correct Rate (WCR) is defined in this paper:

$$WCR = \frac{Count(Correct)}{Count(Total)} \qquad (1)$$

Count(Correct) is the number of words recognized correctly, and *Count(Total)* is the total number of words to be recognized. WCR describes the performance of dialogue system with a certain number of grammar rules. The average *WCR* of the system with four different grammar scales is called *WCR$_a$*. It can be calculated:

$$WCR_a = \frac{\sum_{scale\in SSet} Count(Correct)}{\sum_{scale\in SSet} Count(Total)} \qquad (2)$$
$$SSet = \{500,\ 1000,\ 2000,\ 4000\}$$

The average WCR of twenty evaluators on the system with certain scale grammar rules is called *WCR$_{sca}$,* which can be calculated through the following formula:

$$WCR_{sca} = \frac{\sum_{i=1}^{n} WCR(i)}{n} \qquad (3)$$

The number n is the number of evaluators. There are twenty persons to participate in our experiments.

The evaluation results show that dialogue system has different recognition performance with different grammar sizes. According to Figure 1, the observable trend is that there is a consistent reduction of system performance with increased grammar size.

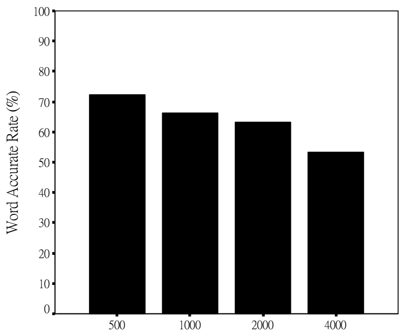

Figure 1: Word accurate rate and grammar size

Figure 1 shows that recognition accuracy drops from 72.4% to 53.4% with a mean of 63.85% when grammar size is increased from 500 to 4000. This observation suggests the need to improve system performance by using dynamically constructed hierarchical grammars instead of monotonic grammars for every recognition slot. Dynamically constructed hierarchical grammars are different from monotonic grammars in that grammar rules are typically classified into several groups according to their prior probabilities to be recognized. The prior probabilities can be obtained from context and other related information. How to get operable hierarchical grammars will be an important part of our future work on RAMCORP.

See Figure 2 for system performance with the 20 subjects.

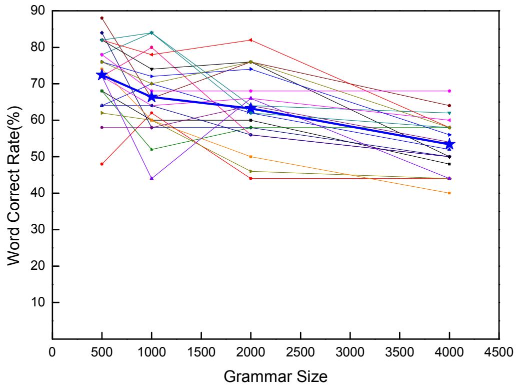

Figure 2: System performance with the 20 subjects in the experiment

There is considerable fluctuation in WCR for the 20 subjects with a standard deviation of 7.56, as demonstrated in Figure 2, which is expected for a telephony dialogue system. It should be noted that the twenty evaluators are non-native English speakers from China so the actual WCR of the evaluated system would be higher than the WCR values required in our experiments if the callers were native speakers requesting the translation of terminologies from English to Chinese.

Figure 3 shows that, across the four grammars on average, the system had varying degrees of performance with the 20 subjects. The maximum is 75.0% and the minimum 49.5% with a mean of 64.1. The standard deviation is 7.56%. Such variations are expected for a telephony dialogue system open to a wide range of speaker diversity.

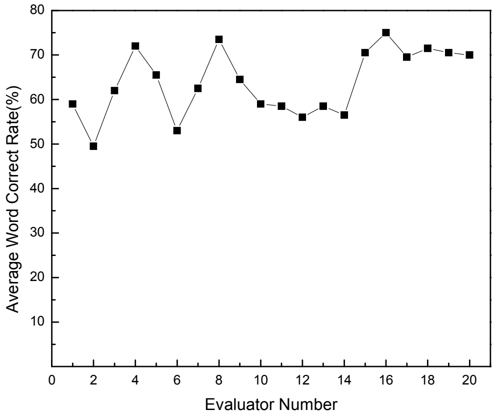

Figure 3: Average WCR with different evaluators

3.3 WCR Variation and Word Length in Characters

Word length defined in number of characters is the second parameter concerned in this study that is expected to have an impact on recognition performance. The WCR based on character length is called WCR_{cl}. It is an average value calculated according to Equation (4):

$$WCR_{cl} = \frac{\sum_{scale\in SSet \ \& \ i\in ESet} Count(Correct_{cl})}{\sum_{scale\in SSet \ \& \ i\in ESet} Count(Total_{cl})} \qquad (4)$$

where, SSet is scale set which represents the same meaning in Equation (3). ESet is the evaluator set {1, 2, 3, ..., 20}. *Count(Correct$_{cl}$)* is the correctly recognized number of words with length “character length (abbr. cl)”. *Count(Total$_{cl}$)* is all test words which length is equal to cl. The evaluation results are summarized in Table 2. The second column in Table 2, marked *Test Set*, lists the word length distribution of all the test words randomly selected in the experiment with **#** indicating the actual number of words selected and **%** its proportion in all of the test words selected. The third column, *Lexicon*, is the distribution of all words in the dictionary with **#** indicating the total number of words of the concerned length and **%** the proportion of such words in the dictionary.

It can be seen that the word length varies from 1 to 21 characters and that the selected words in the test set form a good representation of those in the lexicon in terms of distribution of character lengths. Words with lengths between 4 and 12 characters account for about 90 percent of total number.

C	Test set		Lexicon		WCR$_{cl}$
	#	%	#	%	
1	6	0.15	32	0.06	50.00
2	4	0.10	248	0.46	75.00
3	79	1.98	841	1.56	35.44
4	218	5.45	2399	4.45	53.67
5	320	8.00	3995	7.41	49.06
6	471	11.77	5958	11.05	58.81
7	588	14.70	7187	13.33	61.22
8	528	13.20	7554	14.01	61.36
9	572	14.30	7306	13.55	71.15
10	456	11.40	6066	11.25	72.15
11	313	7.83	4448	8.25	70.93
12	177	4.42	3133	5.81	71.19
13	136	3.40	2043	3.79	75 .74
14	59	1.47	1240	2.30	64.41
15	38	0.95	744	1.38	73.68
16	18	0.45	388	0.72	77.78
17	9	0.22	216	0.40	66.67
18	5	0.13	81	0.15	100.00
19	2	0.05	38	0.07	100.00
21	1	0.03	5	0.01	0.00
M	4000	100.00	53916	100.00	63.85

Table 2: WCR based on character length

As Figure 4 clearly shows, words with different character lengths have different impact on system performance as suggested by *WCR$_{cl}$*. It can be observed from the graph that there are some ups and downs at the two ends of *WCR$_{cl}$-length* curve. This phenomenon can be caused by two possible reasons. Firstly, words shorter than 4 and longer than 12 characters in length are relatively small in population. The randomly selected few cannot support statistic results sufficiently. Secondly, the evaluators involved in these experiments are non-native speakers of English while all the test words were selected randomly from a large dictionary. Therefore there were unfamiliar words for the evaluators, which resulted in inaccurate pronunciations and subsequently recognized inaccuracies by the system.

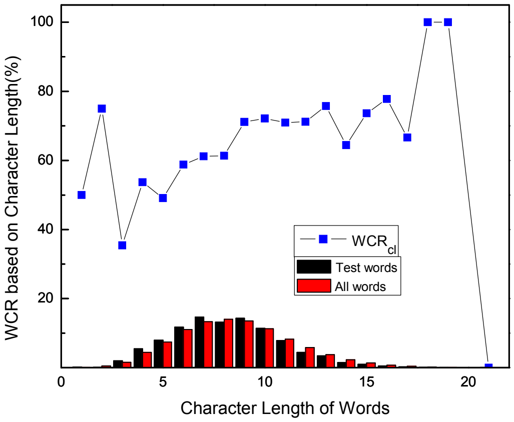

Figure 4: WCR based on character length

But the predominate words with lengths between 4 and 12 have a consistent trend and *WCR$_{cl}$* increases steadily with word length. Generally, the longer a word is, the more likely the word is accurately recognized. One observation is that words between 6 and 8 characters in length have a similar WCR while those between 9 and 12 have a similar but higher WCR. This suggests that the use of word character as a measurement unit has a wide range of variation in terms of WCR, which calls for the use of another measurement unit that exhibits a lower degree of variation. As a result, we introduced the use of syllables as a second measurement unit, to be discussed in 3.4 below.

Based on the evaluation results of *WCR$_{cl}$*, a more suitable dialogue model can be designed for improving performance of dialogue systems. Simple dialogue modules can be applied to recognize long words because these words have a relatively high *WCR$_{cl}$*. Conversely, complex dialogue modules with extended interactive turns will be needed for shorter words that typically have a lower *WCR$_{cl}$*. By doing so, a dialogue system with a good balance between conciseness and accuracy can be achieved.

3.4 WCR Variation and Word Length in Syllable

As mentioned above, words of different lengths have different impact on system performance measured in *WCR*$_{cl}$. In fact, the major factor can be attributed to syllable information, which influences the accuracy of word speech recognition significantly. In this sense, the number of syllables of a word may demonstrate more precisely the correlation between word length and recognition accuracy.

For this purpose, a machine-readable pronunciation dictionary was used to retrieve the number of syllables for each of the test words selected for the experiment. The WCR based on syllable length, *WCR*$_{sl}$, is calculated by the following formula:

$$WCR_{sl} = \frac{\sum\limits_{scale \in SSet \ \& \ i \in ESet} Count(Correct_{sl})}{\sum\limits_{scale \in SSet \ \& \ i \in ESet} Count(Total_{sl})} \qquad (5)$$

The formula is similar to Equation 4. The only difference between them is that *Count(Correct*$_{sl}$*)* is the word count with syllable length "sl" being recognized correctly. The *WCR*$_{sl}$ results are listed in Table 3.

S	**Test set**		**Lexicon**		**WCR$_{sl}$**
	#	**%**	**#**	**%**	
1	429	10.73	4028	0.06	50.00
2	1283	32.07	15582	0.46	75.00
3	1103	27.58	15501	1.56	35.44
4	775	19.38	11020	4.45	53.67
5	293	7.32	5322	7.41	49.06
6	91	2.27	1871	11.05	58.81
7	19	0.47	507	13.33	61.22
8	7	0.18	86	14.01	61.36
M	4000	100.00	53916	13.55	71.15

Table 3: WCR based on syllable length

The first column **S** shows the word length in terms of syllables. The second column in Table 3 is the syllable length distribution of all test words with **#** indicating the actual number of words selected and **%** the proportion of such words in the total number of test words. The third column, marked **Lexicon**, is the distribution of all words in the machine-readable pronunciation dictionary. **#** indicates the actual number of words of a certain length and **%** the proportion of such words in the lexicon. As can be seen from the table, the selected words and the lexicon show good similarity in terms of distribution, suggesting that the test data are sufficiently representative. Words of up to 6 syllables in length make up more than 99 percent of the total test set with a small margin of proportion for words with 7 syllables or more.

Figure 5 is a graphical representation of Table 3. It can be observed that *WCR*$_{sl}$ for words with less than 7 syllables shows a consistent rise as a function of syllable number, increasing steadily together with the increase of word length measured in terms of syllables.

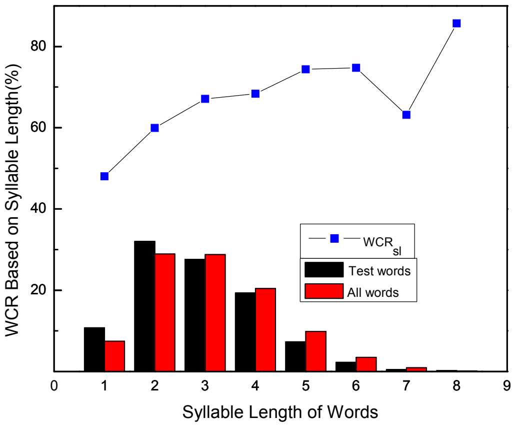

Figure 5: WCR based on syllable length

Compared the results with Figure 4, we can determine that the *WCR*$_{cl}$ jump from 5 characters to characters is because words with 5 characters and 6 characters will have different syllables which influence the accuracy of their speech recognition. A similar phenomenon happens in 8 characters and 9 characters in Figure 4. The evaluation results offer support for designing an effective dialogue system.

4 Conclusions

This paper presented an experiment to evaluate the performance of Nuance for its recognition accuracy measured in word accurate rate (WCR). While conventional measurement is typically conducted in conjunction with grammar size, we designed a novel approach to measure WCR as a function of word length measured in terms of characters and syllables. Results show that while WCR drops with the increase of grammar size, there is also the tendency for WCR to rise as a function of word length. Between characters and syllables, the experiment demonstrated that the latter is a better indication of the correlation between WCR and word length.

The results confirms the conventional wisdom in the first place that, instead of using a monotonic grammar which tends to be large in size and therefore affects WCR, a hierarchical grammar generated dynamically should be preferred for better WCR. This raises an interesting suggestion for the RAMCORP project to augment the list of terminologies in such a way that they can be effectively sub-classified in order to reduce recognition space and therefore to increase WCR. Secondly, the results suggest that better system performance can be expected when RAMCORP moves into a stage that involves the recognition of longer terminological phrases.

The most significant suggestion from the experiment is that a dynamically constructed dialogue model can be possibly achieved based on the word returned by the recognition slot. Such a model can be driven by a probabilistic engine that considers grammar size and word length measured in characters and syllables. Within such a probabilistic dialogue model, modules with different interactive turns can be selected according to the word recognized and returned by the system. While the general principle is that shorter terminologies require more dialogue turns to achieve a completed transaction, the system can be fine tuned for even better transaction completion rate based on probabilities associated to each keyword in the grammar. Such a dialogue system will require a self-maintenance mechanism of the grammar that updates itself for recognition probabilities for each individual rule.

On the basis of the suggestions above, future work will be carried out in two key areas: one is to construct effective hierarchical grammar rules using context and other features of the terminologies concerned in RAMCORP. The other is to design a probabilistic dialogue model for improving the usability of the service through maximally enhanced system performance. In addition, similar evaluation is required for the other languages involved in the project, including Chinese in the first instance and Korean and Japanese in the future.

Acknowledgments

This research was supported by the project "A Remote-Access Multilingual Corpus-Based System for Linguistic Applications (RAMCORP)", City University of Hong Kong. The authors would like to thank Mr John Poon and Ms Lanying Cheng of Nuance Communications, Hong Kong. The authors would also like to thank all the evaluators for their kind help with the experiments reported in this article.

References

Dybkjr, L. and N.O. Bernsen. 2001. Usability evaluation in spoken language dialogue systems. In *Proceedings of the workshop on Evaluation for Language and Dialogue Systems,* volume 9. pp 1–10.

Hartikainen, M., E. Salonen, and M. Turunen. 2004. Subjective evaluation of spoken dialogue systems using servqual method. In *Proceedings of ICSLP.* pp 2273–2276.

Higashinaka, R., N. Miyazaki, M. Nakano, and K. Aikawa. 2004. Evaluating discourse understanding in spoken dialogue systems. *ACM Transactions on Speech and Language Processing,* 1:120.

Hildebrandt, B., H. Rautenstrauch, and G. Sagerer. 1996. Evaluation of spoken language understanding and dialogue systems. In *Proc. ICSLP,* volume 2. pp 685– 688.

Larsen, L. B. 2003. Issues in the evaluation of spoken dialogue systems using objective and subjective measures. In *Automatic Speech Recognition and Understanding.*

Nanjo, H. and T. Kawahara. 2005. A new ASR evaluation measure and minimum bayes-risk decoding for open-domain speech understanding. In *IEEE ICASSP.* pp 1053–1056.

Park, W., S.H. Han, Y.S. Park, J. Park, and H. Yang. 2007. A framework for evaluating the usability of spoken language dialog systems (sldss). *Usability and Internationalization.* pp 398–404.

Walker, M.A. C.A. Kamm, and D.J. Litman. 2001. Towards developing general models of usability with paradise. *Natural Language Engineering,* (6). pp 363–377.

Walker, M.A., D.J. Litman, C.A. Kamm, and A. Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In Philip R. Cohen and Wolfgang Wahlster, editors, *Proceedings of the 35th Annual Meeting of the ACL*. pp 271–280.