LANGUAGE MODELING FOR CONTENT EXTRACTION IN HUMAN-COMPUTER DIALOGUES

Wolfgang Reichl, Bob Carpenter, Jennifer Chu-Carroll, Wu Chou

Lucent Technologies Bell Laboratories,
600 Mountain Avenue, Murray Hill, NJ 07974

ABSTRACT

In this paper we discuss the role of language modeling in a novel natural language dialogue system designed to automatically route incoming customer calls. We arrive at two significant conclusions: First, standard word error rate measures do not reflect application specific requirements; highly reliable content extraction is possible with relatively high word error rates. Secondly blending human-human data with human-machine data did not improve the performance in language modeling.

1. INTRODUCTION

The first task facing any call center is call routing, which involves directing customers to the appropriate branch of the call center for handling their particular request. Call centers currently direct customers calls using touch-tone based interactive voice-response (IVR) system and/or human operators. Our goal is to provide a spoken dialogue-based automated call routing system with performance approaching or exceeding that of human operators in terms of both routing accuracy and naturalness of the dialogues. We restricted our attention to systems for which the routing behavior could be trained with minimal human intervention based on transcriptions of human-human routing dialogues.

In this paper, we focus on the speech recognition and understanding component of the system. In particular, we will discuss methods for acquiring the dictionary, acoustic models, statistical language models, and content models from training data. We arrive at two significant conclusions. First, the standard word error rate measure of recognition accuracy provides a poor indicator of content extraction. Second, the addition of human-human dialogue data or text data to human-machine data impairs speech recognition rather than improving it.

2. CALL ROUTING SYSTEM

Typical IVR systems often frustrate callers due to the rigid hierarchical nature of their menu systems. When responding to a human operator's query of “How may I direct your call?", a caller typically provides either the name of the destination to which they would like to be transferred, or a short description of what activity they would like to perform. Our application involves a large financial services call center handling business covering customer accounts, insurance, loans and financial planning adding up to thousands of activities. Learning these associations is the primary source of difficulty for human operators performing call routing, especially given that requests for some activities, such as loans, are relevant to several destinations depending on the particular activity being performed, such as applying for a loan or making a payment. With 4497 transcribed training calls, we found 23 destinations with at least 10 training instances. Requests by department name accounted for 21.1% of the calls, whereas requests by activity accounted for 72.7% of the calls. The remaining 6.2% of the calls were so long and indirect that they are filtered and sent to a human operator.

Our approach to routing is based on an information retrieval paradigm. During training, we collect transcriptions of customer calls and sort them by the destination to which they should be routed. This task is non-trivial because most call centers do not have well defined specifications of routing behavior and human operators make a significant number of routing errors. We then filter the transcribed text through the morphological component of Bell Labs Text-to-Speech system [10] in order to extract the morphological roots of words. For instance, singulars, plurals and gerunds are reduced to their roots, as are the various verb forms. For instance, “service”, “servicing” and “services” are all represented by the root “service”, and “deposit”, “depositing”, “deposits” are all reduced to “deposit”. At run time, this is represented as a mapping from surface forms to underlying root forms. Forming equivalence classes helps overcome the data sparseness problem because data with similar routing behavior are clustered. Although we focus on equivalence classes generated by morphological roots, in principle other mappings could also be used. For instance, “mortgage" and “home loan" would have been natural candidates for equivalency in our domain, as would have “car”, “automobile”, and “Buick”.

After extracting the morphological roots of the terms, we remove words representing noise in the input, such as “uh” and “um”. Next, we remove function words and other words that are irrelevant for routing purposes, such as “the” and “want”. The words to ignore are the so-called “stop words'” common in information retrieval applications. We employ a slightly expanded version of the list supplied with the SMART system [11]. We then look for sequences of roots uninterrupted by stop words occurring with a frequency above a threshold. In our case, we required single terms (unigrams) to occur at least three times and pairs (bigrams) and triples (trigrams) to occur at least twice. We did not find any sequences of four terms occurring more than twice. From 4497 training calls we extracted 420 unigram terms, 275 content term bigrams, and 62 trigrams of content terms. It was clear that the lexical acquisition process had not yet converged after 4497 training dialogues.

User queries are processed by the speech recognizer to yield a first-best word sequence hypothesis. This is then filtered and morphologically reduced in the same way as the training data. Routing is performed by vector-based information retrieval, as described in [6]. Perhaps more interestingly, clarification subdialogues for cases of vague or ambiguous caller queries are also generated automatically using a novel application of information retrieval techniques [6].

The only work on call routing of which we are aware is that by Gorin et al. [4], who designed an automated system to route calls to specialized AT&T Operators. They select salient phrase fragments from caller requests, such as “made a long distance”. After extracting phrase fragments, they compute likely destinations by either computing a posteriori probabilities for routing or by passing the weighted fragments through a neural network classifier. They then propose hand-built dialogue systems in order to disambiguate. One interesting contrast between our domain and theirs that makes direct comparison of results impossible is that even though they only worked over 14 destinations, their destinations were much more confusible than ours in terms of their terminology and callers were much more likely to make requests that require the attention of two destinations.

3. LANGUAGE MODEL DEVELOPMENT

The automatic speech recognition system for the call router is implemented in a flexible client/server architecture [7] to handle spontaneous caller requests over the telephone network. Two major issues to be considered for the call routing task are robustness and real-time performance. Therefore a one-pass beam-search recognizer, capable of handling a trigram language model is used [8]. The acoustic modeling is based on triphones obtained from a phonetic decision tree clustering [5]. This approach provides the possibility to construct a complete set of triphones necessary for task-independent acoustic training. We used about 80 hours of telephone speech training data, including Switchboard, OGI spontaneous telephone conversations and 25,000 utterances of general English phrases recorded at Bell Labs. The ASR feature vector is a 38-dimension mel-lpc based cepstral vector, without energy component. The tied-state HMMs have about 4000 distinct states and a total of 30,0000 Gaussian mixture components. To compensate for different channel characteristics we implemented a real-time cepstral mean normalization procedure.

The ASR system uses a trigram based language model estimated with Katz’s backoff and Good-Turing discounting [1]. For the purpose of training language models, about 30,000 calls were collected and transcribed. The first 15,080 utterances are recordings of customer/operator (human/human) interactions, while the second set of recordings consist of 15,707 simulated human/machine interactions collected during an internal field trial. The average length of caller utterances in response to human operators is about 10.5 words. Many of the callers start their sentences with a personal greeting and then state the intended activities rather than simply providing the desired destinations. Two typical examples of these calls are listed in Table 1.

Corpus	Call type	Examples
C1	Customer/ Operator	hi valerie I need to ask somebody a question about my checking account
		yes ma’am I’m trying to find someone in deposit services
C2	Customer/ Machine	I’d like to speak to deposit services
		credit card services please

Table 1: Customer/Operator vs. Customer/Machine examples

In the field trial customers knew they were talking to an (simulated) automatic system and their requests were much more direct. Many callers directly named the desired destination or service. The average sentence length for these calls is about 4.4. The distribution of the call lengths for the different corpora are printed in Figure 1.

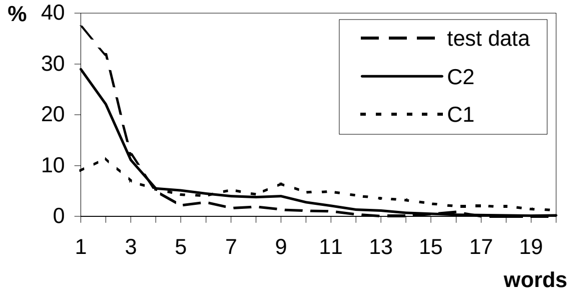

Figure 1: Call lengths by corpora

The vocabulary for call routing was selected after excluding singletons and proper names from the two corpora. About 2,100 words were adopted as vocabulary and they cover most of the collected utterances.

Several approaches of generating language models were studied and evaluated in the end-to-end process of recognition and understanding. First a large out-of-domain language model from the North American Business News task was used for transcription purposes. The idea was to use a generic ASR system to save the long and costly task of transcribing all utterances manually and then using these transcriptions for language modeling. Due to the poor performance of such a generic ASR system in the mismatched acoustic conditions of real-world recordings this was not practicable and the effort to use automatically derived transcriptions was dropped.

We then used the manually transcribed data to generate different language models. One language model was estimated from the customer/operator calls, another from the customer/machine simulations. Since both data sets were collected within the same domain we believed they are both useful for language model training. A third LM was estimated from the union of both sets. The properties of these models are listed in Table 2.

Data	customer/ operator	customer/ machine	both
No. Bigrams	27,200	12,808	32,699
No. Trigrams	68,456	25,664	85,395
Bigram PPX	105.8	32.1	38.2
Trigram PPX	99.5	24.4	29.1

Table 2: N-gram counts and perplexity of languages models

The language models were evaluated with 1,120 sentences collected from real customer calls. A prototype version of the call router was deployed on a few lines of the call center, and the data recorded and transcribed. The average sentence length in these calls was 2.7 and most of them are very focused and short. Informal observation indicates that many callers were simply surprised by the machine answering and responded to the machine's prompt with very short queries. About 1.7% of the words found in the live trial were not in the vocabulary, most of them proper names (e.g. some callers asked to be transferred to a particular person) and truncated words. The bigram and trigram test set perplexities (PPX) for these utterances are listed in Table 2 for the different language models. Surprisingly, the best results were obtained on the smallest language model trained purely on human/machine data. The language models obtained from the human/human conversations result in a high test set perplexity. This can be mainly attributed to the different style and length of the conversations between customer/operator and customer/ machine. The human/machine training data is closest to the test sentences and results in the best language model estimates. Even blending human/human and human/machine training sets increases perplexity by about 20%.

3. ASR EVALUATION

The role of the automatic speech recognition is often taking to be that of supplying an understanding module with a transcription of the user's utterance. But not all errors are equal. Some errors make it impossible to extract the intended meaning of the utterance, while others have very little effect on the semantic content. Some work has been published on the evaluation of voice enabled telecommunication services [2,3], but most of the results are application and system dependent.

Our evaluation was carried out on 2,913 customer calls. These sentences were recorded in adverse but realistic acoustic conditions with plenty of noise, background speech, hum and speech cutoffs. We decided not to clean the data and to report results for realistic conditions. The word based results for LM1 (customer/operator and customer/machine training data) and for LM2 (customer/ machine data only) are listed in Table 3. By all measures, LM2 outperforms LM1. This shows again how important it is to use task-specific training data for the language model estimation. Differences in the speaking style between training and test even within the same task degrade performance. The total word error rate of about 30% for LM2 includes about 6% insertions. The adverse acoustic environment and many hesitations in naturally spoken dialogues often introduce insertions of short function words in the recognizer output, which usually do not degrade routing performance. Standard information-retrieval measures of recall (percentage of correct words that were found) and precision (percentage of words found that were correct) are also listed in Table 3.

	LM1 (blended)	LM2 (hum/mach)
Word accuracy	68.7	69.9
Word correct	74.9	75.8
Precision / Recall	77.4 / 75.5	78.6 / 76.0

Table 3: Word Accuracy and Precision/Recall in %

Because the test utterances were only loosely transcribed, several errors in Table 3 are due to transcription errors (e.g. missing plural "s"). Furthermore, there are many errors related to missing or substituted filled pauses ("um" vs. "uh"). The following shows a typical example of recognition:

recognized:
"I want check the balance of **um** savings **accounts**"
reference:
"I want **to** check the balance of **uh my** savings **account**"

The calculation of the standard word accuracy for this sentence results in 64% words correct (7 out of 11 correct), although the meaning of the sentence is perfectly preserved. Towards an end-to-end evaluation of the call routing system we are looking for an error measure that better reflects the different effects of recognition errors on the subsequent processing steps.

The information extraction of the call routing is based on an information retrieval paradigm and is sensitive only to N-gram terms of content words. In extracting these N-gram terms, function words are removed and salient content words are reduced to their uninflected root form. The reference and recognized sentence from the previous example will be translated to:

terms recognized:
"check balance savings account"
terms reference:
"check balance **my** savings account"

The content word (term) error rate for this example is 80%, which more appropriately reflects recognition performance. This sentence would be correctly routed, despite the missing "my". In Table 4 we present the term accuracy and precision/recall for the N-gram terms. The term accuracy for both language models is about 85%. The term error rate, considering only salient content words in uninflected root form, is only half of the word based error rate. The unigram precision/recall corresponds to the word based precision/recall in Table 3 but is about 16/12 percent points higher. Almost all of the found trigram terms are correct (98.5%), while about 84% of the occurring trigrams were detected. It is interesting to note that the differences between the two language models in Table 4 are smaller than in Table 3. Many errors arise in the recognition of function words and have little effect on content term extraction.

	LM1 (blended)	LM2 (hum/mach)
Term accuracy	84.8	85.0
Term correct	87.1	87.5
Unigram Precision / Recall	94.1 / 87.9	93.7 / 88.4
Bigram Precision / Recall	96.9 / 85.4	96.5 / 85.5
Trigram Precision / Recall	98.5 / 84.3	98.5 / 83.6

Table 4: Content Term Accuracy and Precision/Recall in %

The term error rate is application specific simply because the definition of salient content words depends on the task. It represents a quality measure of the speech recognition system based on factors that determine end-to-end performance. As content was defined for our routing system, this measure provided more encouraging results, in part because content term extraction is highly robust in the face of the levels of noise found in realistic applications. Our content-based recognition metrics also support the evaluation of different recognizer configurations (e.g. different acoustic and language models) on the overall application.

One reason our performance on content term extraction was so high is that by definition the content terms occurred above a given frequency threshold in the training data. But it is important to keep in mind that our language models were word based rather than content-term based, so the language model did not benefit from any clustering based on morphological roots. Furthermore, it is crucial that we derive good models for the “garbage" surrounding our content words for two reasons. First, the context of the content words is important in detecting them. Without modeling “the” properly, we would not be able to estimate the likelihood of it being followed by “balance” and thus the likelihood of “balance" occurring at all.

Overall routing performance is calculated as percentage of successfully routed calls. But some of these calls are ambiguous and our system generates multiple possible destinations as an input to a dialogue-based disambiguation system [6]. Our system routed 10% of the incoming calls to a human operator, and of the remaining calls correctly routed 89% of the remaining calls. The performance on manually generated transcriptions as opposed to recognition results boosted performance to 94%. The difference indicates the overall effect of ASR recognition errors. A detailed analysis of the routing errors is presented in [9].

4. SUMMARY

We described and evaluated the speech recognition component of a domain independent, automatically trained call router. In so doing, we learned two important lessons. First, human/human data is of limited utility in developing language models for human/machine dialogues. People simply speak differently to machines than they do to humans. This is reflected in the dramatic reduction in perplexity in trigram models from 99.5 from customer/operator data, to 24.4 for customer/machine data, and 29.1 from a blend of the two, despite the fact that we had almost twice as much customer/operator data as customer/machine data.

The second lesson we learned is that content is easier to extract than exact transcriptions of customer utterances. In particular, we recovered task relevant content term roots at a much higher rate of accuracy (88.4% recall / 93.7% precision) than would have been expected from the raw word accuracy (76.0% recall / 78.6% precision). Even more surprisingly, for bigrams and trigrams of content terms, precision increases and recall decreases only slightly when compared to individual term extraction (85.5% recall / 96.5% precision for bigrams; and 83.6% recall / 98.5% precision for trigrams).

5. REFERENCES

[1] Katz, S. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer, IEEE Trans. on Acoustic, Speech and Sign. Proc., Vol. 35, No. 3, pp. 400-401, March 1987.

[2] Aust, H., Ney, H., Evaluation Dialog Systems Used in the Real World, ICASSP 98, pp. 1053-1056, Seattle, 1998.

[3] Narayanan, S., Subramaniam, M., Stern, B. Hollister, B. Lin, C., Probing the relationship between qualitative and quantitative performance measures for voice-enabled telecommunication services, ICASSP 98, Seattle, 1998.

[4] Gorin, A., Riccardi G., Wright. J., "How may I help you?", Speech Communication, Vol. 23, pp. 113-127, 1997.

[5] Reichl, W., Chou, W., Decision Tree State Tying Based on Segmental Clustering for Acoustic Modeling, ICASSP 98, pp. 801-804, Seattle, 1998.

[6] Chu-Carroll, J., Carpenter, B., Dialogue Management in Vector-Based Call Routing, Proc. ACL and COLING, pp. 256-262, Montreal, 1998.

[7] Zhou, Q., Lee, C-H., Chou, W., Pargellis, A., Speech Technology Integration and Research Platform: A System Study, EUROSPEECH 97, Rhodes, 1997.

[8] Zhou, Q., Chou, W., An Approach to Continuous Speech recognition Based on Layered Self-Adjusting Decoding Graph, ICASSP 97, pp. 1779-1782, Munich, 1997.

[9] Carpenter, B., Chu-Carroll, J., Natural Language Call Routing: A Robust Self-Organizing Approach, ICSLP 98, to appear, Sydney, 1998.

[10] Sproat, R. (editor), Multilingual Text-to-Speech Synthesis: The Bell Labs Approach, Kluwer Academic Publishers, 1998.

[11] Salton, G., The SMART Retrieval System, Prentice Hall, 1971