PORTABILITY CHALLENGES IN DEVELOPING INTERACTIVE DIALOGUE SYSTEMS

Yuqing Gao, Liang Gu, and Hong-Kwang Jeff Kuo

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598

ABSTRACT

Statistical methods commonly used in developing interactive dialogue systems require large amounts of training data to achieve high accuracy and robustness. This becomes a major bottleneck in building free-style dialogue systems in a new domain or for a new language. Portability challenges hence arise regarding how to build statistical models rapidly and with low cost in terms of data collection, transcription and annotation. In this paper, we discuss challenges as well as potential solutions in several critical issues of efficient language modeling, utilization of untranscribed speech data, automatic annotation, and cross-lingual modeling. We believe that current approaches in these areas are far from mature and call for serious efforts from the research community.

1. INTRODUCTION

We face many common research challenges in developing interactive dialogue systems (IDS), whether they are monolingual human-machine dialogue systems or bilingual machine-mediated human-human dialogue systems (such as speech-to-speech translation systems [1,2]), although each has specific unique difficulties. In this paper, we discuss the common difficulties and our suggestions for solutions to them.

Systems that can handle all forms of spoken dialogue are still an AI-complete problem. When various approximations have been made, dialogue systems can be classified into two kinds. One is known as system initiative or directed dialogue, another is mixed initiative. The system-initiative system usually asks the user a set of questions and expects simple direct answers. These applications are easier to build and an effort to standardize dialogue management is underway, e.g. the VoiceXML standard with W3C. The system coverage is typically determined by hand-written grammars.

In contrast, for mixed-initiative dialogues, the user or system can take charge of the dialogue at any time in offering new information or a new topic. Statistical language models (LMs) for automatic speech recognition (ASR), speech understanding and the dialogue manager become necessary. LMs accommodate the variety of expressions that the user may say spontaneously, which is hard to cover by pre-designed grammars. A statistical natural language understanding (NLU) parser is needed to extract the important semantic entities from the spoken utterances. A more sophisticated dialogue manager is also needed to control the dialogue flow and generate system prompts, etc. Two-way free form speech-to-speech translation also falls under this framework, since the conversations are typically open-ended. Another class of applications that have become technologically viable is natural language call routing, where the system tries to determine the general intent of a caller in response to an open-ended system prompt such as "How may I direct your call?" or "What can I do for you?" With open ended questions, the types of responses by users are even less predictable. For these types of dialogues, which are more natural and efficient for the human user than directed dialogues, challenging issues arise regarding not only how to build the best statistical models (such as for the LM and NLU), but also how to build them rapidly and with low cost in terms of data collection and processing, including transcription and annotation.

Challenges in IDS research can be broadly categorized into two issues: system performance and system portability. System performance includes accuracy, noise robustness, system response time, and issues related to the user interface, all of them directly affecting end-user satisfaction. Although there are still spontaneous user utterances that are difficult to handle, such as, "I'd like to fly from San Francisco to Newark on the twentieth, oh, no, make that next Friday and JFK and from there I'd like to come back the following Wednesday," we are able to build interactive dialogue systems with relatively high accuracy for self-service interaction and transaction applications with a reasonable user satisfaction rate, because the cumulative efforts in the human language technologies areas over the last 20-30 years have paid off in improving system performance [3].

On the other hand, system portability challenges are related to how flexible new IDS for new applications or new languages can be built quickly at a reasonable cost. The success of such systems in the marketplace is necessary to justify further research in industry and to some extent in academia as well. It is well known that the performance of statistical models improves with more training data, leading to the famous quote "There is no data like more data." However, we are starting to question the veracity of this adage in our research by challenging ourselves

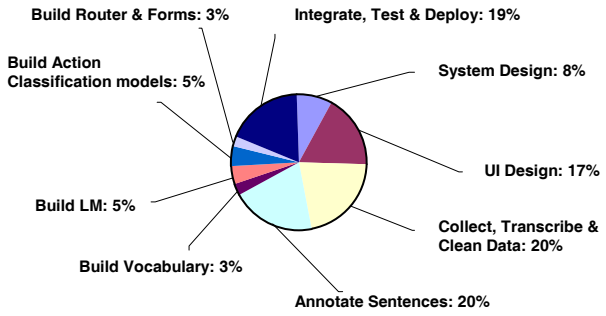

Figure 1. Relative cost of various elements in an IDS development cycle

with the interesting scientific question of how to use less data to get the same performance through new algorithms such as active learning. In this paper we will focus on the portability challenges with the hope this will help foster more research in both industry and academia to develop technologies that will drastically reduce the development cycle and costs of building new natural spoken dialogue applications so that IDS will become successful and ubiquitous. This type of research has previously not received much attention, but is currently in urgent demand.

It is obvious that signal processing and acoustic models are channel and platform sensitive, but are less sensitive or are invariant for different applications. There has been effort in developing language-independent ASR and more specifically acoustic models [4] to reduce the ASR development cost of new languages. The most serious bottlenecks for rapidly developing dialogue applications is from the need for application specific data for the training of statistical models, such as language models for ASR, semantic parser for natural language understanding, classification models for call routing, and domain knowledge for dialogue management.

Specifically, we examine Figure 1, which shows the relative costs of different parts of the development cycle. The major time and cost to build a new application includes developing an understanding of the business logic and application requirements, collecting, transcribing, and cleaning speech data, and using the consensus business logic to annotate sentences or assign action class labels. In order to enhance the IDS portability, one has to simplify or reduce the cost for these processes. Practical problems of portability and cost reduction can be solved by novel technical approaches and scientific and systematic solutions, so they call for serious research by the entire community, rather than just the speech industry. In this paper, we list some specific problems and discuss examples of solutions or research directions.

2. CHALLENGES FOR PORTABILITY

We focus on the statistical language and understanding models used in interactive dialogue systems and how to efficiently train them with much less data or cost than conventional methods. Some of the on-going challenges we face include the following. How do we bootstrap models with little or no data? How do we use existing models and adapt them using a little domain-specific data? How do we reduce the initial cost of transcription and annotation? How can we take advantage of large amounts of data that are collected once the system is put out into the field? Can we take advantage of untranscribed and un-annotated data? Or can we find a more intelligent way to transcribe/annotate them (e.g. active learning)? How do we ensure transcription or annotation consistency and correctness? How do we deal with multi-lingual issues, e.g. porting to new languages, especially for resource-poor languages?

3. APPROACHES

In this section, we describe some of the approaches to improve portability, i.e., given some prior experience, data, and models, what can we re-use and what additional data do we have to collect? After collecting the data, what solutions are needed to develop a new application with reduced cost? Specifically, we discuss four areas of research that are relevant to our goals to reduce data requirements and development costs for new applications. The first area involves efficient statistical language modeling to reduce the amount of domain-specific data that must be collected, for example, by Wizard-of-Oz. Secondly, we consider how to take advantage of large amounts of untranscribed speech data for language modeling. Then we discuss the issue of reducing the cost of annotation or labeling data that are to be used in training either the semantic parser or categorical call classifier. Finally, we touch briefly on cross-lingual language modeling, an issue that still requires a lot of additional research.

3.1. Efficient Statistical Language Modeling

Statistical language modeling (LM) of the probability of various word sequences is crucial for high-performance ASR of free-style open-ended dialogues. During the past decades, much effort has been devoted to better estimate the word sequence probability distributions. The approaches for LM aim at either improving model accuracy using existing large training corpora (such as Broadcast News or Switchboard), or adapting (or porting) these language models to specific domains with a very limited amount of training data [5]. In this paper, we will focus on the latter goal (i.e. model portability). Current approaches to enhance LM portability fall into three categories: 1) obtaining additional training material; 2) interpolating domain-specific LMs with other LMs; 3) improving distribution estimation robustness and accuracy with limited in-domain resources.

Automatic data collection and expansion is the most straight-forward way to achieve an efficient LM, especially when little or no training data is available. Most approaches retrieve additional data from the World Wide Web (WWW), while differing in 1) www search query generation; 2) filtering out only the relevant text from the retrieved pages [6, 7, 8]. LM can be improved by just using web-based n-gram counts [9]. Conversational style data may be further retrieved by adopting the most frequently occurring tri-grams in the conversational Switchboard corpus as search queries [10]. A more sophisticated query generation was proposed in [11] that gradually create relevant queries from the most relevant to the least. The searched sentences were filtered and selected based on an N-gram based similarity measure. In addition, existing conversational corpora in other domains could be used. In Figure 2, we show results of recent experiments demonstrating the ability of the new algorithm to achieve the same WER performance as a baseline system trained on 4-10 times more in-domain data. A huge gain is achieved when there is little training data, for example less than 1,000 sentences (on the left side of the graph). Note that our goal is different from prior work that uses Web data to augment LMs that are already trained on a large amount of data. We are interested in achieving good performance with very little in-domain data in order to quickly build new dialogue systems.

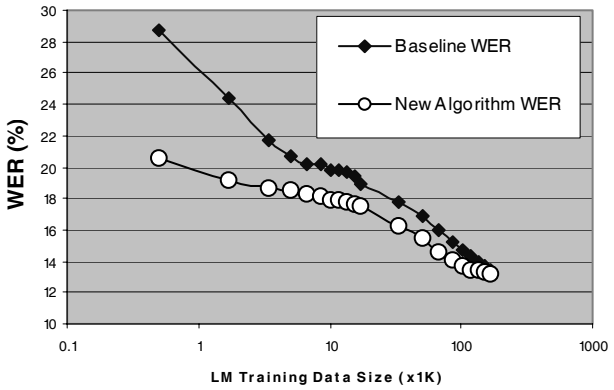

Figure 2. Performance comparison between baseline SLM and improved SLM on real-time interactive dialog system

The LMs trained using various resources can be combined using interpolation-based methods, including Maximum-Likelihood-based model merging and smoothing [12], dynamic cache modeling [13] and MAP-based adaptation.

The third line of research on LM uses topics to enhance both model accuracy and domain coverage. In [13], training data is partitioned into topic-dependent clusters to build mixture language models via the EM algorithm. In [14], a clustering-classification strategy was proposed that performs topic detection during LM decoding and then uses the topic-adapted LMs trained on topic-clustered training data. These two approaches are actually two special cases of a homogeneous mixture language modeling framework that defines and builds LMs based on homogeneity.

3.2. Language Model Adaptation in the Presence of A Large Amount of Untranscribed Speech Data

After a spoken dialogue system is put into the field, and real callers start to use the system, new speech data begin to arrive at a rate that is faster than can be manually transcribed. Since the language usage patterns of real callers may be quite different from what was collected prior to field deployment, especially if system prompts have changed, it is desirable to adapt the language model to the new data.

However, human transcription of speech data is costly, takes a lot of time, and is prone to errors. A few strategies to overcome this problem exist. One method is to use the speech data in a completely unsupervised way [15, 16]. The speech data can first be decoded to create word lattices, from which confidence measures can be derived, which are used in the LM adaptation so that high confidence words have bigger effect than low confidence words [17]. Another method, known as active learning, relies on detecting a small subset of sentences that may be most useful for adapting the system and are thus specially selected to be transcribed by humans. In one study [18], active learning was shown to reduce by 27% the amount of transcription needed to achieve a particular word accuracy. Combining active and unsupervised learning can reduce the word error rate by 75% in another study [19]. In a different experiment [20], using unsupervised data for LM adaptation was shown to be quite effective (19% improvement) versus the baseline, but still lagged the performance when using human transcribed data. Finally, we also note that there is new research being pursued to utilize untranscribed speech data to directly optimize the performance of a natural language call routing system [21].

3.3. Semi-Automatic Annotation for Training Understanding Models

To build a speech understanding model (whether it is for a NLU parser or a call classifier), one typically needs labeled data for training. For the NLU parser, the annotated data are often in the form of parse trees in a Treebank, for example, and for the call classifier, simply the class/destination/action labels.

A variety of methods have been explored to deal with data sparseness and the human costs of annotation, as well as the inconsistency in labeling among labelers. In [22], a hybrid rule-based and statistical parser is proposed using recursive weighted finite state transducers. Simple rules can be defined which are used in an initial parse of the corpus. Corrections are done manually and the statistical models are then re-trained based on the corpus. In [23, 24], various unsupervised and active learning approaches are proposed for training categorical classifiers for natural language call routing.

As Young described in [3], for semantic understanding, although a flat parse tree model [25] can be adapted to work with relatively simple training data annotations (e.g. an unaligned list of semantic tags), the representational power of the flat model is not generally adequate. On the other hand, the hierarchical HUM [26] requires fully annotated tree-bank data. Any attempt to use simpler annotations and let EM discover the hidden structure are very unlikely to work since there are far too many degrees of freedom in an unrestricted context free model.

One can develop algorithms to further simplify the semantic annotation process, and combine multiple parsing results, thus improving the manual annotation efficiency [27]. The approach in [27] casts the problem of semantic annotation as a classification problem: each word is assigned a unique set of semantic tag(s) and/or label(s). This is in contrast to the conventional parsing strategy which is designed to generate a complete parse tree for a sentence. The method enables ``local'' semantic annotation resulting in partially annotated sentences and hence minimizing human labor. It also proposed to use SVM and similarity-based classifiers to produce multiple parsing results. These two algorithms both outperform the baseline decision-tree parser when there is little data.

Based on these results, a tool was also developed to combine the three parser outputs in order to improve the parsing accuracy using less data, and detect correctly parsed sentences and thereby reduce the annotation cost. The proposed method reduces the annotation time and cost significantly as shown in Figures 3. As a result, to achieve 80% accuracy, the amount of data required dropped from 6.3K sentences to 2K sentences, a reduction of more than 3 times the amount of data needed. For 85% accuracy, the data needed dropped from 9K sentences to 5K sentences, representing a reduction of about 1.8 times, as shown in Fig 3b. Furthermore, the percentage of sentences that need to be human corrected is reduced by about 2 times (from 100% manual correction rate to 50%, as shown in Figure 3a).

3.4. Cross-lingual Language Modeling

For resource poor languages, a large amount of domain-specific text for statistical language model estimation would be very difficult to obtain. The techniques proposed in [28], which

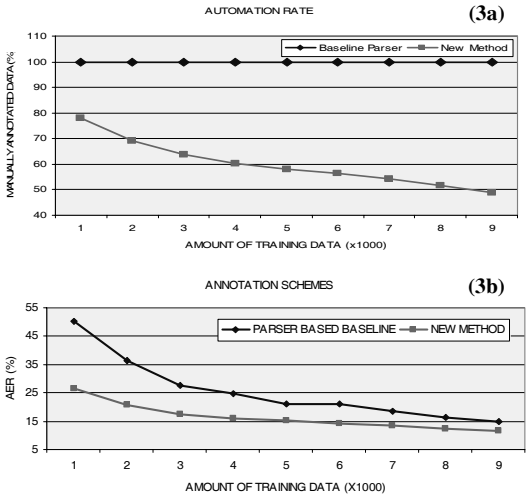

Figure 3. Reduction of Manual Annotation Rate (3a) and improvement of Annotation Accuracy (3b) using methods presented in [27].

exploit domain specific text in a resource-rich language to adapt a language model in a resource-deficient language, open an interesting research area for rapid development of LMs for resource-poor languages. Similar ideas in this area are worth further investigation. A primary advantage of the technique described in [28] is that in the process of cross-lingual language model adaptation, it does not rely on the availability of any machine translation capability. It uses ideas from cross lingual latent semantic analysis to develop a single low-dimensional representation shared by words and documents in both languages. Future research is needed before the approach can be applied to build a LM for a completely new colloquial language, e.g., Pashto, since the approach needs a parallel corpus to derive a dictionary, and uses a unigram model only.

4. CONCLUSIONS

Statistical methods have long been the dominant approach in speech recognition and more recently have also been extended to other areas of spoken dialogue systems. Although the power of statistical methods has made ASR a mature technology [4] to some extent, the nature of statistical and probabilistic models requires large amounts of training data, which should be matched with the application condition to achieve favorable performance. This has limited the flexibility and portability of the approaches and the development of dialogue systems using these approaches. One major research challenge is therefore to develop new algorithms to reduce the amount of data needed to train high performance statistical models.

In this paper we share our experience in exploring various approaches to increase the model portability to new applications and new languages. We also provided a brief survey of some approaches that are intended to reduce the development cycle and cost. Current results range from 2.5 to 10 times reduction in the amount of data needed for LM training, and 2-3 times reduction in the amount of human annotation effort for training the NLU parser. There is a wealth of other promising approaches such as hybrid statistical and knowledge (rule-based) systems that use rules to overcome data sparseness. Still to be studied is how resources such as linguistic databases and higher level information can be advantageously utilized. We believe portability challenges should be a new research area that has important practical ramifications, and call for serious research from the community, especially from academia.

5. REFERENCES

[1] Y. Gao, et al, "MARS: A statistical semantic parsing and generation-based multilingual automatic translation system," Machine Translation, vol.17, pp.185-212, 2002.
[2] L. Gu, F.-H. Liu, Y. Gao and M. Picheny, "Improving Statistical Natural Concept Generation in Interlingua-based Speech-to-Speech Translation," in Proc. Eurospeech, 2003.
[3] S. Young, "Talking to Machines (Statistically Speaking)," in Proc. ICSLP, 2002.
[4] T. Schultz, "Towards Rapid Language Portability of Speech Processing Systems," in Proc. Conf. on Speech and Language Systems for Human Communication, 2004.
[5] J. Bellegarda, "Statistical language model adaptation: review and perspectives," Speech Communication, vol. 42, pp. 93-108, 2004.
[6] A. Berger et al, "Just-in-time language modeling," in Proc. ICASSP, pp. II:705-708, 1998.
[7] S. Schwarm, I. Bulyko and M. Ostendorf, "Adaptive language modeling with varied sources to cover new vocabulary items," IEEE Trans. Speech and Audio Processing, vol. 12, no. 3, 2004.
[8] V. B. Le, et al, "Using the web for fast language model construction in minority languages," in Proc. Eurospeech'2003.
[9] X. Zhu and R. Rosenfeld, "Improving trigram language modeling with the world wide web," in Proc. ICASSP, pp.I:533-536, 2001.
[10] I. Bulyko, M. Ostendorf and A. Stolcke, "Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures," in Proc. HLT-NAACL 2003.
[11] R. Sarikaya, A. Gravano and Y. Gao, "Rapid language model development using external resources for new spoken dialogue domains," in Proc. ICASSP'2005.
[12] R. Iyer, M. Ostendorf and H. Gish, "Using out-of-domain data to improve in-domain language models," IEEE Signal Processing Letters, vol. 4, no. 8, pp. 221-223, 1997.
[13] R. Iyer and M. Ostendorf, "Modeling long distance dependence in language: Topic mixtures versus dynamic cache models," IEEE Trans. Speech and Audio Processing, vol. 7, pp.30-39, 1999.
[14] P.-C. Chang and L.-S. Lee, "Improved language model adaptation using existing and derived external resources," in Proc. ASRU'2003.
[15] A. Stolcke, "Error Modeling and Unsupervised Language Modeling," in Proc. of the 2001 NIST Large Vocabulary Conversational Speech Recognition Workshop, 2001.
[16] M. Bacchiani and B. Roark, "Unsupervised Language Model Adaptation," in Proc. ICASSP'2003.
[17] R. Gretter and G. Riccardi, "On-line Learning of Language Models with Word Error Probability Distributions," in Proc. ICAASP'2001.
[18] D. Hakkani-Tur, G. Riccardi, and A. Gorin, "Active Learning for Automatic Speech Recognition," in Proc. ICASSP'2002.
[19] G. Riccardi and D. Hakkani-Tür, "Active and Unsupervised Learning for Automatic Speech Recognition," in Proc. Eurospeech'2003
[20] K. Visweswariah, et al "Task Adaptation of Acoustic and Language Models Based on Large Quantities of Data," in Proc. ICSLP 2004.
[21] V. Goel, H.-K. J. Kuo, S. Deligne, W. Cheng, "Language Model Estimation for Optimizing End-to-End Performance of a Natural Language call Routing System," in Proc. ICASSP'2005.
[22] A. Potamianos and H.-K. J. Kuo, "Statistical Recursive Finite State Machine Parsing for Speech Understanding," in Proc. ICSLP'2000.
[23] G. Tür, M. Rahim, and D. Hakkani-Tür, "Active Labeling for Spoken Language Understanding," in Proc. Eurospeech'2003.
[24] G. Tür and D. Hakkani-Tür, "Exploiting Unlabeled Utterances for Spoken Language Understanding," in Proc. Eurospeech'2003.
[25] R. Pieraccini et al, "Stochastic representation of semantic structure for speech understanding," Speech Communication, Vol. 11, no.2, 1992.
[26] S. Miller, et al, "Statistical language processing using hidden understanding models," in Proc. HLT Workshop, 1994
[27] R. Sarikaya, et al, "Fast Semi-Automatic Semantic Annotation for Spoken Dialog Systems", in Proc. Interspeech-2004.
[28] W. Kim & S. Khudanpur, "Cross-Lingual Latent Semantic Analysis for Language Modeling", in Proc. ICASSP'2004.