Applying Natural Speech to Domain-Specific Human-Machine Dialogue

Noam Lotner, Michal Gishri, Dikla Shechnik, Ami Moyal, Vered Aharonson
Afeka Center for Language Processing, Afeka Academic College of Engineering

I. Abstract

The ultimate goal of human-machine interaction is natural spoken dialogue that flows as freely as the human-human communication we are all accustomed to. For this interaction to occur, the machine must be able to, not only recognize the words of the human speaker, but to interpret their meaning well enough to provide a relevant and useful response. However, today's technological reality cannot yet handle these tasks to a satisfactory level. In this paper we show that, limiting the dialog to a specific domain and using a combination of speech recognition and textual processing, may enable such dialogue. The beginning stages of the research, presented here, focus on a limited question-answering system. The system uses ASR (Automatic Speech Recognition) technologies to produce a textual equivalent of the user's input speech and then applies a distance measure to match to the appropriate response from an answer bank.

II. Introduction

This paper presents ongoing research on a Human-Machine dialogue system in American English. The domain selected for the research is the 1990-1991 Persian Gulf War. This particular domain was selected for several reasons. On the one hand, the Persian Gulf War is an historical event that can be defined by a beginning and an end, thus eliminating the possibility of continuous expansion of the domain. This means that the word list (and thus transcription lexicon) can be defined in advance and will remain unchanged. On the other hand, the topic is broad enough that the vocabulary is still larger than what current dialogue systems support, and allows the speaker relative freedom in the use of natural speech without limitations on the structure of the input speech.

The challenges of developing a dialogue system are endless. In addition to the challenges associated with textual dialogue systems, which include parsing issues, disambiguation and semantic analysis, etc. a speech-based system must also deal with partial input, resulting from standard speech recognition errors such as insertion, deletion and substitution. Unique to this research, is the integration of text-based algorithms with algorithms used for speech analysis. This integration has the potential to take Human-Machine dialogue a step further.

III. Approach

Our focus in this research is to combine textual search and matching of user questions to system responses, together with in-depth analysis of the speech recognition results in a way that will enable us to reach better performance. Textual analysis of the domain allows us to identify the relevant keywords that will enable us to direct the speech recognition for best understanding the user's request. On the other hand, familiarity with the weaknesses of speech recognition can help predict probable errors and assist the textual search in supplying useful responses to the user. Combining these two into a single distance-measure (between the input and the system responses) will allow us to efficiently search for the most useful information to present to the user.

Figure 1 illustrates the flow of the system.

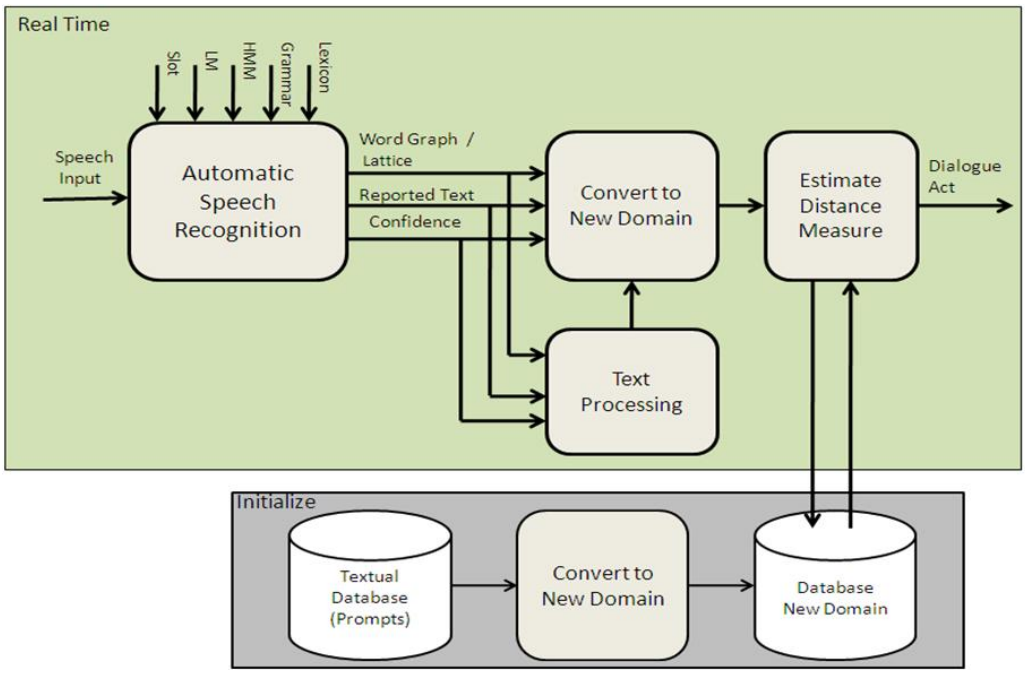

Figure 1

A pre-processing stage creates a textual, domain-specific database that contains the system's responses together with the various tag information. Once the user initiates the dialog (poses a question, in this case), the speech is fed into an ASR engine to produce the most probable words spoken. This list of words is then matched with the textual database, using the distance-measure to produce a list of several responses which best match the user's request.

IV. Experimental Framework and Results

The first stage of the research was to build the infrastructure for the dialogue system. To this effect, a textual database on the Gulf War domain was compiled and a word list was extracted. The second stage included experiments with textual algorithms for defining the criteria that will lead to optimal machine comprehension in the best case scenario – where the user input is not dependent on the speech recognition performance (i.e. 100% recognition). Textual input was used in order to simulate this condition. In order to begin this process, a Question-Answering system was built, where the user asks a question and the system uses a distance measure to select the best matched response from the fully indexed database. The distance measure was formulated based on the following linguistic criteria:

- **Weight** - each word was provided a value between 1 and 5 (determined by objective criteria).
- **Lexeme grouping** - words with a common lemma are algorithmically linked, i.e. start, started (based on the British National Corpus high frequency word list and adapted to American English).
- **Semantic grouping** - words with similar meanings are algorithmically linked, i.e. start, begin.
- **WH tagging** – where responses are tagged by the type of WH question they respond to.

Each of these criteria was tested separately and in combination in order to evaluate their individual and combined effect on the distance measure and their overall contribution to attaining the best results.

In the final stage we integrated an ASR engine based on the HTK open-source toolkit, and the Festival text-to-speech software to create a fully

interactive system based on speech input and output.

The acoustic models we used for speech recognition were trained on the Macrophone database (Telephone channel recordings of 4005 speakers, for a total of 44 hours of speech). The feature-set we used was the standard 13 MFCC with 1st and 2nd derivatives. We used the ARPA phoneme-set consisting of 39 phonemes, each modeled as a 3-state HMM with left-to-right transitions. Six additional models were used for the various types of background noise that exist in the database. These were then clustered into 3383 context-dependent triphone states. Output probabilities were modeled using mixtures of 16 Gaussians per state. We used a bigram language model that was trained on the domain-specific textual data that was collected, with a lexicon size of 25K words.

The test speech used for this experiment consists of 4 US English speakers that were each recorded uttering the 122 questions that were defined in the previous stage.

Table 1 shows the results of the text-only version using the various linguistic criteria. The test set consists of the 122 questions, with 120 responses in knowledge-base:

Table 1

N-best list size	Word matching	+KW Weighting	+Lemma Grouping	+Semantic Grouping	+WH-tag
1	80.33%	79.51%	87.70%	87.70%	91.80%
2	86.07%	86.89%	94.26%	95.08%	95.08%
3	88.52%	92.62%	95.08%	95.90%	95.90%
4	89.34%	94.26%	95.90%	96.72%	96.72%
5	89.34%	94.26%	96.72%	97.54%	97.54%

In order to indentify the impact of a larger database of responses, the indexed data was expanded to include 1710 responses. Table 2 shows the test results, using the combined score:

Table 2

N Best	Performance
1	91.80%
2	95.08%
3	95.08%
4	95.09%
5	96.72%

Table 3 shows the results of using the speech input. The textual database consisted of the original 120 responses.

Table 3

N Best	3Performance
1	75.90%
2	84.21%
3	85.60%
4	86.43%
5	88.64%

V. Discussion and Future Work

As can be inferred from our results, it is possible to achieve relatively good results when limiting the dialog to a specific domain, and using a combination of text processing and speech recognition. The test set we used is still too small to enable a full dialog, but preliminary results show that the textual part of the system can be expanded with little impact on response relevance. It remains to be seen whether this remains true when using speech input. Other

topics that need further investigation are: in-depth analysis of the typical mistakes made by the speech-recognition errors and incorporating the information into the response-matching score, further refinement of the textual distance measure (including implementation of keyword tags), and improving the speech models by using richer data sets.

VI. Acknowledgments

This research was funded the Israeli Ministry of Defense, Mafat.

VII. References

Bernstein, J., Taussig, K., and Godfrey, J. (1994). *Macrophone: an American English telephone speech corpus*. LDC94S21. CD-ROM. Philadelphia: Linguistic Data Consortium.

Bertomeu, N., Uszkoreit, H., Frank, A., Krieger, H.-U., and Jörg, B. (2006). *Contextual phenomena and thematic relations in database QA dialogues: results from a Wizard-of-Oz experiment*. In Proc. of the HLT-NAACL 2006 Workshop on Interactive Question Answering.

Folch, H., and Habert, B., (2000). *Constructing a Navigable Topic Map by Inductive Semantic Acquisition Methods*, Extreme Markup Languages 2000

Gishri, M., Silber-Varod, V. and Moyal, A., (2010). *Lexicon Design for Transcription of Spontaneous Voice Messages*. LREC2010 (May 17-23) Malta.

Kyoung-Soo Han, Do-Sang Yoon, Joo-Young Lee, and Hae-Chang Rim. (2004). *A Practical QA System in Restricted Domains. In Workshop on Question Answering in Restricted Domains*. 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004), pages 39–45, Barcelona, Spain.

Lison, P. and Kruijff, G. J. M. (2009). *Efficient parsing of spoken inputs for human-robot interaction*. (2009). The 18th IEEE International Symposium on: Robot and Human Interactive Communication, pp. 885 - 890.

Shechnik, D. and Moyal, A., (2010). *An Efficient Algorithm for Vocabulary Reduction*. IEEE 2010 Eilat, Israel, November 17, 2010.

Shen D. and Lapata M. (2007) *Using semantic roles to improve Question Answering*. In Proceedings of EMNLP-CoNLL,2007.