Current challenges in dialogue systems
GSLT-course “Speech Technology”
Maria Eskevich

Abstract

Dialogue is the natural part of normal human behavior. We do talk to organize our life, our work, our future. Nowadays technical progress imposes the other integral part – computer and interaction with it. Computers are not only calculators anymore. They can help our civilization not only to discover the world, the space, but to make our life easier. Since it seems easier to say something to the computer that rules your house while doing some other work instead of doing this work or even type it.

This paper describes current challenges in dialogue systems research and application. We have to understand that these two types of results might be diverse because of the different purposes: to verify how good the new technology is, how much more innovative the proposed methods are or how efficient is its application.

Introduction

For the long period the conversational computers were the reality of scientific fiction, because scientists needed to achieve evident improvement in closely related domains – like speech technologies, language processing, dialogue modeling. And they surely needed more effective computers to support these technologies.

We have to be aware of the hierarchy of speech applications which requires special technologies:

Voice dictation systems which provide only transcription of what the user dictates to the system;

Command-and-control applications or interactive voice response systems which enable users to perform commands with voice input, i.e. they are specializing in isolated word recognition;

Spoken dialogues systems are an advanced type of applications. It’s an interface between the user and the computer-based application which permits spoken interaction with the application in a relatively natural manner, though it always operates in a constrained domain. The active role of the system in these dialogues may vary (from machine-directed dialogue to mixed-initiative strategy, or it might be a combination of these methods, when constraining dialogue is used in case of problems understanding the user).

Fundamental challenges existent in spontaneous speech

As human-machine interaction tends to be similar to human-human interaction scientists drew their attention to the latter. The large corpora were collected and analyzed, since the study of human-human interaction is a special area of research itself.

It is possible to derive following fundamental properties of spontaneous speech which represent the supreme challenges for dialogue systems.

Disfluencies such as filled pauses, repetitions, repairs, and false starts are frequent in natural conversation. Before they were regarded as errors, but more and more they are treated and understood as part of natural conversation. They cause problems for higher level natural language processing and degrade transcript readability for humans. In some contexts modeling disfluencies can also improve speech recognition performance. Most programs are usually taught on a written text, and then the strategy is to detect the interruption point using lexical and prosodic cues (which is not a trivial problem itself) and remove them before further processing. Statistical model based on tree adjoining grammar to represent the possibility of a word being disfluent. Or another approach is a transformation-based learning.

In many formal languages punctuation helps to organize text, to process it and the readability in general. Even modeling sentence-level punctuation can also improve speech recognition performance itself. But for processing of natural conversation, finding sentence boundaries by machine is a challenge. Pauses are neither necessary nor sufficient indicators of sentence boundaries. There are different types of computational models for finding sentences, such as combination of N-gram language models and prosodic classifiers, knowledge sources combined with HMM framework. The important application of punctuation detection for online human-computer dialogue system is a so-called "endpointing", i.e. when dialogue system has to determine when speaker has completed an utterance. It requires the basic disambiguation between hesitation and grammatical pauses. The challenge: "endpointing is an online task and must be accomplished using information only before the potential endpoint in consideration". Usually there is an assumption about the predetermined duration threshold. Prosodic features such as intonation add confidence to true ends before a speaker pauses.

The other type of problems arises when more than one speaker is involved. They do not speak sequentially as in a written pattern of dialogues. It is interesting that familiarity with the other talker does not appear to affect the rate of overlap.

The last point to investigate, to deal with and to use in evaluation is the extra linguistic information. Detecting affect through speech requires more than words. Here we can face various problems. It's hard to obtain emotional data (usually it is acted speech), it's hard to scale emotions, and gestures' data are also very specific to interpret.

Problems with speech recognition

2 main recognitions errors arise from the facts that conditions differ from training conditions and that people might use words unknown to the system.

Some systems proceed straight away from the fact, that it is not possible to use the robust speech recognition techniques – the desirable robust technology is not available on the market or in principle. Then pragmatic approach is brought to the forefront. Such principles as limited mixed-initiative interaction, dealing with dialogue repairs during the dialogue and minimization of the duration of the dialogues are applied.

So, to solve the problem of unknown words some authors propose a speech understanding model to identify and interpret them. When it is not possible to recognize the word, the system tries to determine the semantic category of the misrecognized word. And this information helps the dialogue management system to choose the best dialogue strategy to be applied.

The system might confuse the word with most similar sounding word, which will lead to the semantic ambiguity, or it might manage to label the out-of-vocabulary words and interpret the utterance keeping in mind this feature. It means that the understanding model is based on stochastic representation of semantic and pragmatic knowledge using conceptual segments. A conceptual segment is a word sequence corresponding to the basic units of meaning. There are 3 types of them: referential representing the application domain, illocutionary referring to the speech act theory and filler (words that are irrelevant for the meaning representation).

When we consider such typical problems of error categorization as insertion, deletion and substitution and the utility of the word for the meaning of the utterance we can identify 8 error cases.

Insertion of useless word – it fits into conceptual segment *filler*. When it's just the repetition of the word it means for the system that it stays in the same state of the hidden Markov model.

Insertion of a useful word: at the moment it is impossible to treat the case when it is totally unexpected.

Deletion of a useless word – it is like a skipped state in the hidden Markov Model.

Deletion of a useful word: it is impossible to completely interpret an utterance when a useful word is deleted, but still it is possible to try identify the category of deleted word.

Substitution of a useless word by another useless word – the conceptual segment remains the same for the case (filler).

Substitution of a useful word by a useless and of a useful word by useful (it is the hardest point especially when misrecognition within the same word class)

As we define out-of-vocabulary word as a word not included in the word lexicon, for the actual state-of-art in the domain the clue information is the expectation of information. And the system informs the user about misrecognition or if it is not important, i.e. it would not disrupt the interpretation of the rest of the utterance, it just drop it.

Prosody

The use of prosodic information for speech recognition might provide more natural sounding output and assist recognition. Several studies have shown that vocal expression of emotions can be recognized more or less reliably in the case of simulated produced by trained speakers or actors.

Prosody classifier can be based on artificial neural networks and might be combined with a discrete Hidden Markov model for gesture analysis to recognize user internal state of mind in multi-modal applications.

Multimodal dialogue system

It is interesting to examine GEMINI (generic environment for multimodal interactive natural interfaces) project to get insight of multimodal approaches and challenges for a dialogue application.

Its' concept core consists in using modality and language independent representation of a dialogue description. This approach provides a capability to generate an application in several modalities and languages.

The objectives might be described as development of a flexible platform able to produce user-friendly, high quality, multi-modal and multilingual dialogue interface to a wide area of databases with a reduction of human effort to entering parameters while being guided by a graphical user interface and effectiveness and adaptability to new languages.

Many modalities (speech, gesture and facial expressions) are processed at the same moment and are affecting the dialogue and error handling. The system uses many tricks to be more friendly, like incremental help on no input or no match, when at each stage system provides more information in order not to annoy the client. Distinction between novice and expert users – dependent on this information. Implicit and explicit verification.

However, it seems to be more feasible and practical to explore dynamic features of the gestures, namely their sudden changes, pacing, direction, velocity or the acceleration curves as in the system SmartKom. It can be done with the help of Virtual Touch Screen (infrared camera and projector). A proper modeling of the dynamics of the user's gestures is crucial, hidden Markov models can be used to train and classify the gestures in a way similar to speech recognition. Gestures can be conceived as passing some atomic states, so it is possible to classify them into some categories. Since for the experiment were taken different people, speaker-dependent gestures, so there are some misrecognition.

Fusion of different modalities seems to be necessary because user might not use all modalities at the same time

Dialogue system with World Wide Web integration

World Wide Web can give us not only the up-to-date information, but some current or dynamic data. This advantage of the internet together with the lack of keyboard (for example, in the car) and the necessity to carry out other activities (like driving) caused the appearance of such an interesting and still challenging application as in-car based dialogue system that allows guided surfing the internet using speech.

There are some features of this system in particular due to the conditions existent when application is used. The dialogue has to be intuitive and easy to follow since the driver is in a situation when a good pat of is concentration has to cope with the traffic. It's not only the task of information retrieval and information extraction, but the important part is the presentation (computer has to avoid its' superfluity).

Since dialogue system integrated with remote database is a complicated complex consisting of speech processing applications and properly data transferring technologies we have to deal with connection interruptions.

Conclusions

It is obvious that dialogue systems are of interest for researchers and for concerned companies. Though they will always pursue the different purposes (the former will put the emphasis on developing more advanced systems and on testing theories on dialogues, while the latter will concentrate on producing systems that will work efficiently in the real world) the results and new kind of errors they get will bring new insight in speech technology, human-computer communication and language inner system understanding. As the technologies are developing at a very quick rate, it seems that complicated systems using multimodality approaches and web integration have great prospects.

References

Bousquet-Vernhettes C., Vigouroux N. Recognition error handling by the speech understanding. System to improve spoken dialogue systems. *ITRW on error handling in spoken dialogue systems, Switzerland, 2003, p. 113-118*

Clark H.H. When to start speaking, when to stop, and how. *ITRW on error handling in spoken dialogue systems, Switzerland, 2003, p. 1-4.*

Glass J.R. Challenges for spoken dialogue systems

McTear M.F. Spoken dialogue Technology: Enabling the Conversational user interface. *ACM Computing Surveys, Vol.34, No.1, March 2002, p. 90-169*

Shi R.P., Adelhardt J., Zeissler V., Batliner A., Frank C., Noeth E., Niemann H. Using speech and gesture to explore user states in multimodal dialogue systems. *AVSP 2003 – International conference on audio-visual speech processing, France, 2003, p. 151-158*

Shrinberg E. Spontaneous speech: How people really talk and why engineers should care. *Eurospeech 2005*

Schneider R., Heisterkam P. Connecting dialogue systems to the World Wide Web. ITRW on multi-modal dialogue in mobile environments, Kloster Irsee, Germany, 2002, p. 49-55

Seydoux F., Trutnev A., Rajman M. Dialogue management with weak speech recognition: a pragmatic approach. *ITRW on error handling in spoken dialogue systems, Switzerland, 2003, p. 133 – 139.*

Wang Yu-F.H., Hamerich S.W., Schless V. Multi-modal and modality specific error handling in the GEMINI project. *ITRW on error handling in spoken dialogue systems, Switzerland, 2003, p. 139 – 144.*