Spoken language understanding and interaction: machine learning for human-like conversational systems

Milica Gašić

University of Cambridge

Dilek Hakkani-Tür

Google

Asli Celikyilmaz

Microsoft Research

Abstract

In recent years, the interest in research in speech understanding and spoken interaction has soared due to the emergence of virtual personal assistants. However, whilst the ability of these agents to recognise conversational speech is maturing rapidly, their ability to understand and interact is still limited. At the same time we have witnessed the development of the number of models based on machine learning that made a huge impact on spoken language understanding accuracies and the interaction quality overall. This special issue brings together a number of articles that tackle different aspects of spoken language understanding and interaction: clarifications in dialogues, adaptation to different domains, semantic tagging and error handling. These studies all have a common purpose of building human-like conversational systems.

1. Introduction

The emergence of virtual personal assistants is generating increasing interest in research in speech understanding and spoken interactions with machines. However, whilst the ability of these agents to recognise conversational speech is maturing rapidly, their ability to understand and interact is still limited to a

few specific domains, such as weather information, local businesses, and some simple chit-chat. Their conversational capabilities are not necessarily apparent to users. Interaction typically depends on handcrafted scripts and is often guided by simple commands. Deployed dialogue models do not fully make use of the large amount of data that these agents generate. Promising approaches that involve statistical models, big data analysis, representation of knowledge (hierarchical, relations, etc.), utilising and enriching semantic graphs with natural language components, multi-modality, etc. are being explored in multiple communities, such as natural language processing (NLP), speech processing, machine learning (ML), and information retrieval. However, to have natural conversational interactions with these agents, there are still many issues that need to be solved. The aim of this special issue, therefore, is to present new and emerging topics in machine learning which might lead to richer and natural human-computer interaction.

Obtaining meaning from human natural language is a complex process. The potential range of topics is vast and even well-formed utterances can be syntactically and semantically ambiguous. Spontaneous conversational speech naturally contains grammatical errors, repetitions, disfluencies, partial words, and out of vocabulary words. Conducting intelligent conversations over multiple turns requires maintaining the dialogue state over time, dealing with errors that arise from the speech recogniser, determining an adequate dialogue strategy, estimating the quality of that strategy, and generating natural language responses.

Over the years many different approaches and models have been proposed (e.g. syntactic and semantic analysis of spoken text, hybrid models that use speech processing components as features for semantic analysis, learning representations for spoken text, contextual models, statistical models of dialogue). These methods have drawn inspiration from machine learning solutions e.g. sequence tagging, syntactic parsing, and language modelling, primarily because these tasks can be easily abstracted into machine learning formulations (e.g. structured prediction, dimensionality reduction, regression, classification, supervised or reinforcement learning). These representations have evolved into novel

understanding models based on discriminative methods, Bayesian nonparametrics, neural networks, low rank/spectral techniques, and word/phrase/sentence level embeddings based on deep learning methods.

In dialogue modelling, methods based on partially observable Markov decision processes and reinforcement learning have enabled limited domain dialogue models to be built that are trainable from data, robust to noise, and adaptable to changes in the user or domain. Following success in other areas, neural networks have also been applied to different aspects of dialogue modelling yielding significant improvements. The problem remains however as to how to extend these models to exploit the huge datasets that users of virtual personal assistants generate, and thereby enable the richer and reliable conversation that users expect.

Problems in spoken language understanding and dialogue modelling are particularly appealing to those doing core ML research due to the high-dimensional nature of the spaces involved (both the data and the label spaces), the need to handle noise robustly and the availability of large amounts of unstructured data. But there are many other areas within spoken language understanding and dialogue modelling for conversational systems where the ML community is less involved and which remain relatively unexplored, such as semantics, open-domain dialogue models, multi-modal dialogue input and output, emotion recognition, finding relational structures, discourse and pragmatics analysis, multi-human understanding (meetings) and summarization, and cross lingual understanding. These areas continue to rely on linguistically-motivated but imprecise heuristics which may benefit from new machine learning approaches.

2. Overview of the issue

The first paper our our issue ***Modeling the Clarification Potential of Instructions: Predicting Clarification Requests and other Reactions*** by Benotti and Blakbourn investigates clarifications in conversation from a theoretical perspective. They hypothesize that implicatures are a rich source of

clarification requests and motivate this hypothesis in theoretical, practical and empirical terms. The present a model of clarification potential by inferring conversational implicature. They then go on to show that much of the inference can be handled using classical AI planning. They conclude that implicature and claricationfit well together and they investigated their interaction by combining theoretical work from pragmatics, practical work from the dialogue system community and empirical evidence from spontaneous dialogues situated in an instruction giving task.

The second paper of our issue ***Dialogue manager domain adaptation using Gaussian process reinforcement learning*** by Gasic et. al. deals with the issue of adaptation in the context of statistical dialogue modelling. Statistical approaches to dialogue modelling offer cheaper development, robust performance and improvement over time of use. Gasic and colleagues investigate what kind of architecture and statistical models are particularly well suited to support adaptation to changes in the dialogue domain and ultimately underpin open domain dialogue systems. They show that a range of methods, including the incorporation of prior knowledge, Bayesian committee machines and multi-agent learning, facilitate extensible and adaptable dialogue systems.

The third paper ***A Framework for Pre-Training Hidden-Unit Conditional Random Fields and its Extension to Long Short Term Memory Networks*** by Kim et. al. describes a novel technique for spoken language understanding. They propose a unsupervised framework for pre-training hidden-unit conditional random fields (HUCRFs) and apply it several natural language tagging tasks. The key idea is that the proposed framework enables learning model parameters in an unsupervised manner without labelled data to initialize HUCRFs prior to supervised training. The technique proved effective in tagging tasks in natural language. They show that this idea could be extended to other learning techniques including deep learning and they applied the proposed technique to long short term memory (LSTM) networks and obtained gains.

Our final paper ***Improving the Understanding of Spoken Referring Expressions through Syntactic-Semantic and Contextual-Phonetic Error-***

Correction by Zukerman and Partovi deals with the problem of detecting speech recognition errors in a conversational system. They offer a mechanism that uses shallow semantic parsing to break up the referring expressions heard by the ASR into labelled semantic segments, which are then used to set up syntactic expectations. They describe a syntactic-semantic error-correction model that decides how to modify the output of the speech recogniser on the basis of the syntactic expectations of its semantic segments. Finally, they propose a contextual-phonetic model that re-ranks the output of a Spoken Language Understanding (SLU) system on the basis of the phonetic similarity between words misheard by the speech recogniser and the contextually-valid candidate interpretations returned by the SLU system.

3. Future directions

The goal of this special issues is to highlight some of the issues and possible solutions to spoken language understanding and interaction problems from both applied and theoretical perspective and to highlight how machine learning can facilitate the new frameworks that can help advance modern conversational systems. Some key questions that future research need to address include but are not limited to representation and optimisation, data, scalability, multilinguality and multi-modality.

Word-vector models have made a huge impact in Natural Language Processing, but their full potential is yet to be utilised in conversational systems. The question that needs answering is how can ML help provide novel representations and models to capture the structure of spoken natural language especially considering spontaneous conversational speech? Also, the nature of the problem requires new and robust inference and optimisation techniques which is why ML research is so important for furthering the field.

In speech and NLP we typically have large amounts of less useful background data and small amounts of very useful in-domain data. Are current ML algorithms sufficient to gracefully deal with this problem? One of the papers

presented in this issue tackles this problem. However the more general question remains: can we harness non-dialogue data to build dialogue models?

While many speech and NLP problems depend mainly on static speech or text corpora, dialogue is unique in that the user provides an opportunity for learning on-line. Which non-intrusive methods can we use to engage the user is such a way that it leads to improvement of the dialogue models?

The scalability is a main obstacle to wider application of many machine learning techniques. ML-based dialogue systems have only tacked limited domains so far and we need frameworks that can scale to large open domains leveraging the semantic web. This "scalability bottlenecks" is inherent in natural language.

In the area of spoken dialogue systems typically one or a hand-full of languages dominate. However, the underlying machine learning methods are language-independent and in principle allow for development of a system in any language provided sufficient training data. Can adaptation methods be developed to build conversational understanding systems for low resource languages without going through rigorous annotation processes?

Multi-modal conversational systems are particularly appealing as they engage a number of senses and have the potential to produce more human-like interaction. Still, due to the large number of design choices that need to be made, they are largely hand-crafted. Investigating how machine learning can improve these systems is a very promising line of research.