the morning paper

a random walk through Computer Science research, by Adrian Colyer

Made delightfully fast by

A survey of available corpora for building data-driven dialogue systems

JUNE 28, 2016 ~ ADRIAN COLYER

A survey of available corpora for building data-driven dialogue systems Serban et al. *2015*

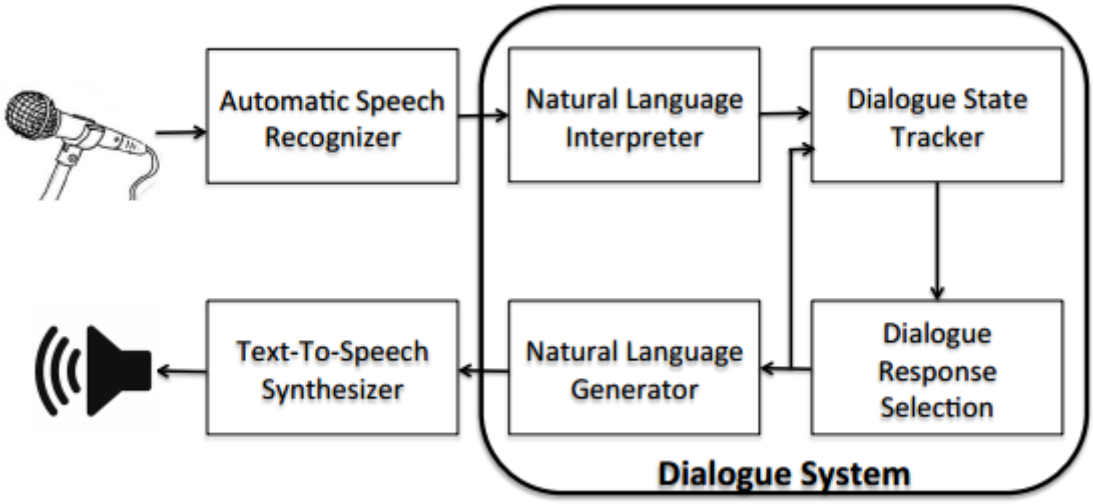
Bear with me, it's more interesting than it sounds :). Yes, this (46-page) paper does include a catalogue of data sets with dialogues from different domains, but it also includes a high level survey of techniques that are used in building dialogue systems (aka chatbots). In particular, it focuses on data-driven systems, i.e. those that incorporate some kind of learning from data.

> … a wide range of data driven machine learning methods have been shown to be effective in natural language processing, including tasks relevant for dialogue such as dialogue policy learning, dialogue state tracking, and natural language generation.

This particular paper is focused on corpus-based learning where you have been able to build up, or have access to, a data set on which you can train your models. If you want to build a defensible machine learning based business, having access to quality sources of data that your competitors don't is a good start. Out of scope is training dialogue systems through live interaction with humans – but there are some references to follow on this so I may well return to that topic later on in this mini-series.

Anatomy of a dialogue system

The standard architecture for a dialogue system looks like this:



Natural language interpretation and generation are core NLP problems with applications well beyond dialogue systems. For building chatbots, where we assume written input and output, the speech recogniser and synthesiser can be left out. I had naively assumed that if you had a good working system that can deal with textual inputs and outputs, it would be a simple matter of bolting a speech-to-text recogniser in front of the system in order to build a voice-driven assistant. It turns out it's not quite as simple as that, since the way we speak and the way we write have important differences:

> The distinction between spoken and written dialogues is important, since the distribution of utterances changes dramatically according to the nature of the interaction.... Spoken dialogues tend to be more colloquial, use shorter words and phrases and are generally less well-formed, as the user is speaking in a train-of-thought manner. Conversely, in written communication, users have the ability to reflect on what they are writing before they send a message. Written dialogues can also contain spelling errors or abbreviations, which are generally not transcribed in spoken dialogues.

Even written dialogue – for example for movies and plays, and in fictional novels – has apparent distinctions from real speech. Which leads to this wonderful observation: "Nevertheless, recent studies have found that spoken language in movies resembles human spoken language." As an occasional movie watcher, I had never thought to question that, or that a study might be necessary to demonstrate it!!

Anyway, I digress. Within dialogue systems we can distinguish between goal driven systems – such as travel assistants or technical support services – where the aim is to accomplish

some goal or task, and non-goal driven systems such as language learning tools or computer game characters. Most startups building chatbots will be building goal driven systems.

> Initial work on goal driven dialogue systems primarily used rule-based systems… with the distinction that machine learning techniques have been heavily used to classify the intention (or need) of the user, as well as to bridge the gap between text and speech. Research in this area started to take off during the mid 90s, when researchers began to formulate dialogue as a sequential decision making problem based on Markov decision processes.

Commercial systems to date are *highly domain specific* and heavily based on *hand-crafted features*. "In particular, *the datasets are usually constrained to a very small task*."

Discriminative models and supervised learning

Discriminative models, which use supervised learning to predict labels, can be used in many parts of a dialogue system. For example, to predict the intent of a user in a dialogue, conditioned on what they have said. Here the intent is the label, and the conditioned utterances are called *conditioning variables* or *inputs*.

> Discriminative models can be similarly applied in all parts of the dialogue system, including speech recognition, natural language understanding, state tracking, and response selection.

One popular approach is to learn a probabilistic model of the labels, another is to use maximum margin classifiers such as support vector machines. Discriminative models may be trained independently and then 'plugged in' to fully deployed dialogue systems.

Answering back

When it comes to choosing what your chatbot is going to say (e.g., in response to a user message) there are again two broad distinctions. The simpler approach is to select deterministically from a fixed set of possible responses (which may of course use parameter substitution):

> The model maps the output of the dialogue tracker or natural language understanding modules together with the dialogue history (e.g. previous

> tracker outputs and previous system actions) and external knowledge (e.g. a database, which can be queried by the system) to a response action.

This approach effectively bypasses the natural language generation part of the system. The fixed responses may have been crafted up-front by the system designers, but there are also systems that effectively search through a database of dialogues and pick the responses from there that have the most similar context:

> …the dialogue history and tracker outputs are usually projected into an Euclidean space (e.g. using TF-IDF bag-of-words representations) and a desirable response region is found (e.g. a point in the Euclidian space). The optimal response is then found by projecting all potential responses into the same Euclidean space, and the response closest to the desirable response region is selected.

More complex chatbots generate their own responses. Using a method known as beam-search they can generate highly probably responses. The approach is similar to that used in the sequence-to-sequence machine translation paper we looked at recently. Short (single request-response) conversations are simpler than those that need to be able to handle multiple interactive turns.

> For example, an interactive system might require steps of clarification from the user before being able to offer pertinent information. Indeed, this is a common scenario: many dialogues between humans, as well as between humans and machines, yield significant ambiguity, which can usually be resolved over the course of the dialogue. This phase is sometimes referred to as the grounding process. To tackle such behaviors, it is crucial to have access to dialogue corpora with long interactions, which include clarifications and confirmations which are ubiquitous in human conversations. The need for such long-term interactions is confirmed by recent empirical results, which show that longer interactions help generate appropriate responses.

Incorporating external knowledge

Chatbots may rely on more than just dialogue corpora for training. When building a goal-drive dialogue system for movies, Dodge et al. ("Evaluating prerequisite qualities for learning end-to-end dialog systems") identify four tasks that such a working dialogue system should be able to perform: question answering, recommendation, question answering with recommendation,

and casual conversation. They use four different subsets of data to train models for these tasks: "a QA dataset from the Open Movie Database (OMDb) of 116k examples with accompanying movie and actor metadata in the form of knowledge triples; a recommendation dataset from MovieLens with 110k users and 1M questions; a combined recommendation and QA dataset with 1M conversations of 6 turns each; and a discussion dataset from Reddit's movie subreddit."

Using external information is usually of great importance to dialogue systems, especially goal driven ones. This could include structured information – such as bus or train timetable for answering questions about public transport – typically contained in relational databases or similar. It's also possible to take advantage of structured external knowledge from general natural language processing databases and tools. Some good sources include:

- WordNet, with lexical relationships between words for over a thousand words,
- VerbNet, with lexical relationships between verbs, and
- FrameNet, which contains 'word senses' for over ten thousand words.

Tools include part-of-speech taggers, word category classifiers, word embedding models, named entity recognition models, semantic role labelling models, semantic similarity models, and sentiment analysis models.

If you're building a new application and don't have ready access to large corpora for training, you may also be able to *transfer* learning from related datasets to bootstrap the learning process. "Indeed, in several branches of machine learning, and in particular in deep learning, the use of related datasets in pre-training the model is an effective method of scaling up to complex environments."

An example of this approach in action is the work of Forgues et al. on *dialogue act classification* (classify a user utterance as one out of *k* dialogue acts).

> They created an utterance-level representation by combining the word embeddings of each word, for example, by summing the word embeddings or taking the maximum w.r.t. each dimension. These utterance-level representations, together with word counts, were then given as inputs to a linear classifier to classify the dialogue acts. Thus, Forgues et al. showed that by leveraging another, substantially larger, corpus they were able to improve performance on their original task.

What were you saying?

Tracking the state of a conversation is a whole sub-genre of its own, which goes by the name of *dialogue state tracking* or DSTC. It is framed as a classification problem: given current input to the dialogue state tracker plus any relevant external knowledge from other sources (e.g. the timetable information from our previous example), the goal is to output a probability distribution over a set of predefined hypotheses, plus a special 'REST' hypothesis which captures the probability that none of the others are correct. For example, the system may believe with high confidence that the user has requested timetable information for the current day. DSTC model include both statistical approaches and hand-crafted systems. "More sophisticated models take a dynamic Bayesian approach by modeling the latent dialogue state and observed tracker outputs in a directed graphical model… Non-bayesian data-driven models have also been proposed."

Longer term memories

We recently looked at Memory Networks and Neural Turing Machines which can store some part of their input in a memory and use this to perform a variety of tasks.

> Although none of these models are explicitly designed to address dialogue problems, the extension by Kumar et al. to Dynamic Memory Networks specifically differentiates between episodic and semantic memory. In this case, the episodic memory is the same as the memory used in the traditional Memory Networks paper which is extracted from the input, while the semantic memory refers to knowledge sources that are fixed for all inputs. The model is shown to work for a variety of NLP tasks, and it is not difficult to envision an application to dialogue utterance generation where the semantic memory is the desired external knowledge source.

Personality

As if all of the above wasn't hard enough, you may also want your bot to exhibit some kind of consistent personality. In fact say the authors, "attaining human-level performance with dialogue agents may well require personalization."

> We see personalization of dialogue systems as an important task, which so far has been mostly untouched.

The Last Word

There's plenty more detail and several additional topics in the original paper, which I have skipped over. If this topic interests you, it's well worth checking out. I'll leave you with the following closing thought:

> There is strong evidence that over the next few years, dialogue research will quickly move towards large-scale data-driven model approaches, in particular in the form of end-to-end trainable systems as is the case for other language-related applications such as speech recognition, machine translation and information retrieval… While in many domains data scarcity poses important challenges, several potential extensions, such as transfer learning and incorporation of external knowledge, may provide scalable solutions.

POSTED IN UNCATEGORIZED

DEEP LEARNING MACHINE LEARNING

< PREVIOUS
On chatbots

NEXT >
A neural conversation model

Multi-domain dialog state tracking using recurrent neural networks | the morning paper (guest)

6 years ago

[…] like those available corpora for building dialog systems might come in […]

Share Vote Reply

End of Term, and the power of compound interest | the morning paper (guest)

6 years ago

[…] A survey of available corpora for building data-driven dialogue systems […]

Share Vote Reply

Data-driven dialog bots | Vcrsoft's Blog (guest)

5 years ago

[…] A survey of available corpora for building data-driven dialogue systems […]

Share Vote Reply

So that was 2016 | the morning paper (guest) 5 years ago

[…] A survey of available corpora for building data-driven dialogue systems […]

Share Vote Reply