Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

MSc Thesis

Learning in Dialogue Interactions

Author:
Dario Chiappetta

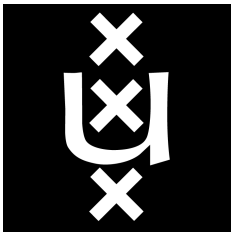
Supervisor:
Prof. Raquel Fernández

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science

in the

Institute for Informatics
Institute for Logic, Language and Computation

May 2013



"Thanks to my solid academic training, today I can write hundreds of words on virtually any topic without possessing a shred of information, which is how I got a good job in journalism."

Dave Barry

UNIVERSITEIT VAN AMSTERDAM

Abstract

Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Institute for Informatics
Institute for Logic, Language and Computation

Master of Science

Learning in Dialogue Interactions

by Dario CHIAPPETTA

A basic fact about human-human dialogue is that there is often more than one way of talking about any domain. For example, instead of saying *"Turn right after 200 meters"*, a route giver may say *"Turn right at Barclays"* given that *"Barklays"* is an enstabilished referring expression. Human speakers efficiently align terminology and associated ontology in talking about any specific domain (in this example we can say that ontologically there is only one entity – a location – but there are different linguistic terms we can use to refer to it). In contrast, current dialogue systems typically have a static ontology and a static vocabulary, which is used in both generation and interpretation, requiring users to formulate their utterances using the terminology known by the system. The goal of this project is to work towards a system that adapts its own linguistic resources (including possible ontology, vocabulary and grammar) to the interlocutor. The system should be able to learn new concepts by assigning new meanings to known words, as well as new words to talk about concepts known by the system in the domain.

Acknowledgements

The acknowledgements and the people to thank go here...

Contents

List of Figures

List of Tables

Abbreviations

AI	**A**rtificial **I**ntelligence
ASR	**A**utomatic **S**peech **R**ecognition
CA	**C**onversational **A**gent
DM	**D**ialogue **M**anager
ISU	**I**nformation **S**tate **U**pdate
MT	**M**achine **T**ranslation
NLP	**N**atural **L**anguage **P**rocessing
SLU	**S**poken **L**anguage **U**nderstanding
TTS	**T**ext **T**o **S**peech

Symbols

c	chunk of text
m	meaning
s	sentence
σ	chunk similarity
τ	word similarity

Introduction

The **user interface**, or human-computer interface (HCI) is the component of a computer system that provides a space of interaction between the human user and the resources offered by the machine; such a space defines a bridge language which human intentions can be translated into, to be converted into computational procedures for the machine; vice versa, the result of the computation is then presented to the user in the same language, which he or she is assumed to understand.

In the early days of computing, the so-called **batch interfaces** were non-interactive: the user/programmer was supposed to feed the machine with a software, punched on cards using the *machine's assembly language* directly, and retrieve the result of the computation printed on paper. The third and fourth generations of computing brought **text-based** interfaces for operating systems (UNIX, DOS, CP/M), that were later took over by **Graphical User Interfaces** (GUI), moving the human-machine interaction on a new, visual language made of windows, icons and buttons. Recently, touch screen and camera devices allowed for the implementation of even more natural means of interaction based on **gestures** and physical actions.

From this brief spot on computing history, and in a way from common sense, we can draw the rather trivial, yet crucial, conclusion that the trend in user interfaces development is to **close the gap** between humans and machines by moving the needle of the interface languages from a machine-centered space towards the human language itself. To this respect, the studies on Natural Language Processing (NLP) assume a dramatically central role, as dramatically central is natural language in the interaction of humans with each other.

Dialogue Systems

A spoken dialogue system, or conversational agent (CA), allow humans and machines to interact through an intermediate language which is as close as possible to the **human language**, and through conversational episodes that implement as close as possible the human dialogue modalities.

Research in dialogue systems has been carried on since the **early days** of Artificial Intelligence. A milestone in the early work on this field is ELIZA Weizenbaum [1966], which provides the user with a basic human-like interaction based on pattern matching; another example is the SHRDLU system , which interfaces the user with a simple spatial domain being able to ambiguous or implicit references to the entities in it.

According to Jokinen and McTear [2009], modern dialogue systems can be divided in **two main types**: task-oriented and nontask-oriented. Intuitively, systems in the first category are meant to deal with a specific task such as making a hotel booking, or booking a plane ticket; an example in this category is the MIT Mercury system, a vocal interface to a flight database Seneff and Polifroni [2000]. On the other hand, nontask-oriented systems are meant to engage in conversations without a specific purpose to fulfill, but the one of delivering a realistic simulation; ELIZA itself is an example of nontask-oriented dialogue system.

Task-oriented systems can be very simple, as simple and well-formalized the task is; many applications, such as travel service or call routing, can be successfully solved by **slot-based** systems: each step of the conversation requires some pieces of information, modeled as slots, to be filled in by the user (departure city, arrival city, date, and so on); given the slots to be filled, the dialogue task can be solved with a formal grammar of interaction. As the complexity increases, more phenomena of human interaction have to be modeled, such as turn-taking, multimodality or grounding , as well as semantic structures such as quantification and negation; slot-based systems are not sufficient to model these scenarios Gabsdil [2003], that require more advanced frameworks such a the Information State Update (ISU) one.

Learning to talk

One of the constituent features of humans' ability to speak is that such an ability is **not innate**, but is rather learned through interactions.

Ever since the power of computers grew enough to allow for intensive statistical analysis of significant amounts of data, **Machine Learning** approaches to Artificial Intelligence tasks got more and more prominent in the scene, often outperforming static methods (i.e. where the solution procedure for a task is explicitly coded by the programmer). One of the clearest examples is the field of NLP itself, where the most important tasks, like parsing or machine translation, are dominated by Machine Learning methods based on corpora, meaning that, for instance, a Machine Translation system will first be trained on a corpus of aligned sentences. Statistical structures will be extracted from this corpus, like the most likely word-by-word alignment, and later be used to process new examples.

However, the task of learning dialogue interactions brings some **peculiar** challenges. First of all, humans learning to talk do not go through two separate phases of learning and processing, but rather improve their abilities episode by episode; as Fernández et al. [2011] point out, this incremental learning structure is nowadays not implemented in state-of-the-art systems. Furthermore, ...

- Learning to produce or understand new surface forms, or **realizations**, for a given meaning – eg. the sentence "Bill eats an apple" for the action of Bill eating an apple.
- Learning to produce new **meanings** from the existing ones and their respective realizations – eg. the concept of motor home, sharing the features of a house and a car.
- Learning a **grammar of conversation**, to place the correct utterance at each step of a conversational episode. – eg. an appropriate answer to the utterance "My name is Bill" is "Nice to meet you", whereas "I like cookies" would not sound as much appropriate.

Lastly, we can point out that the definitions of actions, meaning and episodes are somehow arbitrary and not sharply bounded, as it can be argued for recursive and compositional structures at any level of their interpretation.

This thesis project

The aim of this thesis project is to design and implement language learning capabilities for an existing dialogue system, focusing on the **realization level**. That is, given a fixed list of meanings, the system should be able to classify every sentence into its correct meaning.

The **domain** I am considering is the one of a music player application, being able to get natural language input from the user and translate it into an appropriate corresponding behaviour. For example, when the input is "Play Pictures at an exhibition", the system should start playing the famous suite by Modest Mussorgsky.

In this domain, each **meaning** is an action that a player is performed, and is defined by a set of representative sentences, being its surface forms. For instance, the action of increasing the volume level can be defined by the following set of sentences:

- *Increase the volume*
- *Increase the volume level*
- *Raise the volume*
- *Increase the volume please*

Note that, even though one would be likely to think of such meanings in compositional terms, such compositionality have not been explored for the sake of simplicity.

The **task** for the application is, given an arbitrary input sentence and a context (the point of the conversational episode being realized), to perform the correct action, that is, associate that sentence with its correct meaning. Furthermore, the system should be able to **learn** new realizations for each meaning, as unknown sentences are given in input and processed.

Note that such processing might be more or less **semantically intensive**. As an example, it can be argued that, given the above definition of the action to increase the volume, matching *"raise the volume please"* is an easier task than classifying *"Turn up the volume"*.

Also, an unknown input sentence should be given a **confidence score** for each candidate meaning it is associated to; the system should be able to narrow possible needs for clarification down to single sentence components, eventually asking the user for disambiguation as specifically as possible.

This document is structured as follows. Chapter ...

Appendix A

Appendix Title Here

Write your Appendix content here.

Bibliography

Joseph Weizenbaum. Eliza, a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45, January 1966. ISSN 0001-0782. doi: 10.1145/365153.365168. URL `http://doi.acm.org/10.1145/365153.365168`.

Kristiina Jokinen and Michael F. McTear. *Spoken Dialogue Systems*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2009.

Stephanie Seneff and Joseph Polifroni. Dialogue management in the mercury flight reservation system. In *Proceedings of the ANLP-NAACL 2000 Workshop on Conversational Systems*, ConversationalSys '00, pages 11–16, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1605285.1605288`.

M Gabsdil. Clarification in spoken dialogue systems. In *In AAAI*, 2003.

Raquel Fernández, Staffan Larsson, Robin Cooper, Jonathan Ginzburg, and David Schlangen. Reciprocal learning via dialogue interaction: Challenges and prospects. Proceedings of the IJCAI 2011 Workshop on Agents Learning Interactively from Human Teachers (ALIHT 2011), 2011.