Corpus-driven conversational agents: tools and resources for multimodal dialogue systems development

Maria Di Maro
Department of Human Studies
University of Naples 'Federico II', Italy
maria.dimaro2@unina.it

Abstract

In this paper, it is going to be presented how tools made available through CLARIN can be applied for research purposes in the development of corpus-driven conversational agents. The starting point will be the description of a standard architecture for multimodal dialogue systems. For some of its parts, specific available tools will be briefly described, due to their suitability to their development.

1 Multimodal Dialogue Systems

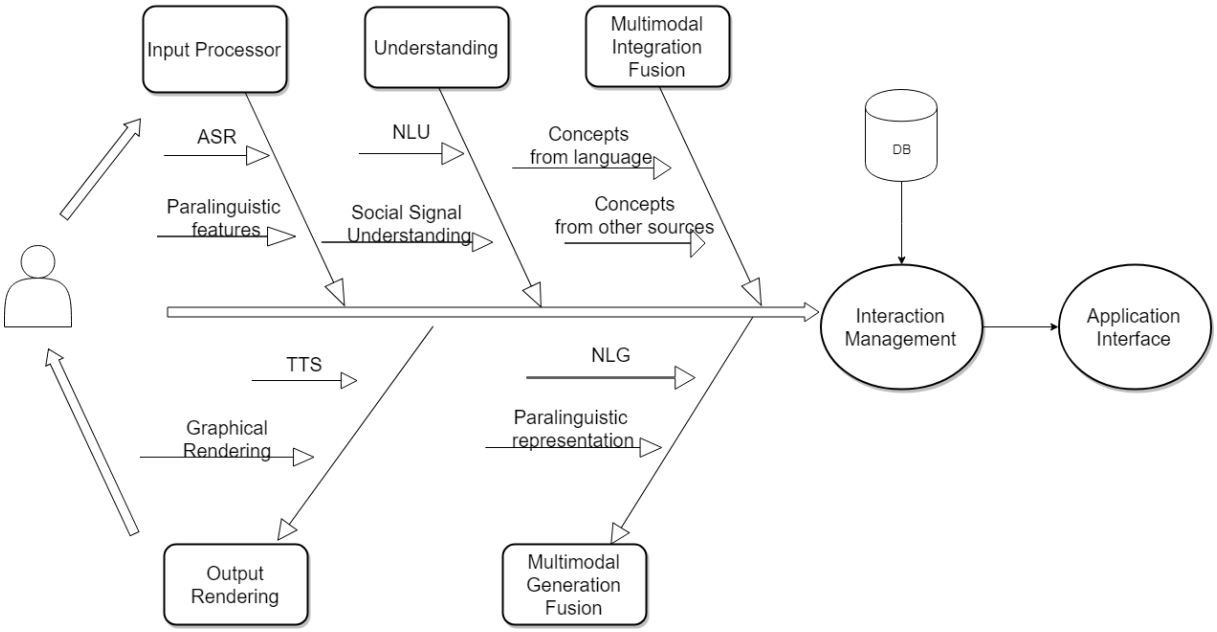
The present paper gives an overview of a PhD research on dialogue modelling for conversational agents, to be framed in the ongoing national project CHROME, whose aim is to define a methodology of collecting, analysing and modelling multimodal data in designing virtual agents serving in museums. The PhD research under consideration is, therefore, mainly concerned with building the interaction with the resulting virtual gatekeeper, which will guide museum visitors in the exploration of cultural contents. In more details, starting from an empirical study of conversational phenomena, especially in cultural heritage domains, common ways of expressing requests and curiosities by visitors, and strategies of communicating cultural contents by guides will be collected and analysed, along with semantic, syntactic and paralinguistic language-dependent strategies.

Conversational Agents are computer systems capable of conversing with humans. These dialogue systems are one of the most currently researched field in Artificial Intelligence, since the ability to communicate one's understanding by means of language is one possible way to manifest intelligence. In the Macmillan Dictionary[1] *intelligence* is defined as 'the ability to understand and think about things, and to gain and use knowledge'. In this definition, one concept draws particular attention: *knowledge*. Building the knowledge base for such systems is the first step to give them *intelligence*. For this particular goal, the use of some tools facilitates the job of interaction designers, such as linguists. In this work, we will concentrate on multimodal dialogue systems, which not only make use of spoken language, but they also use other communication channels to understand and express intents. For this reason, the knowledge to be constructed will comprise different linguistic levels, and also other paralinguistic features to be modelled.

The standard architecture for a multimodal dialogue system consists of different modules, which serves one another to build the interaction (Figure 1). The input elaborated by the user is firstly processed by a first module, which takes the audio produced by the user and transform it in a string to be further analysed. Parallel to that, gestures, facial expression, prosody and other paralinguistic features arising from the interaction are captured by sensors. The classification and consequent understanding of the meaning of the linguistic and paralinguistic inputs is processed in the second module. The meaning associated to the received signals are fused together to recognize a single intent. The decision concerning the flow of the interaction are taken in the Interaction Management module, which is connected to a knowledge base including the information concerning the accomplishment of specific intents. When all

[1] Macmillan Dictionary Online: https://www.macmillandictionary.com/ [last consultation on the 15th June 2018]

Figure 1. Multimodal Dialogue Systems Architecture

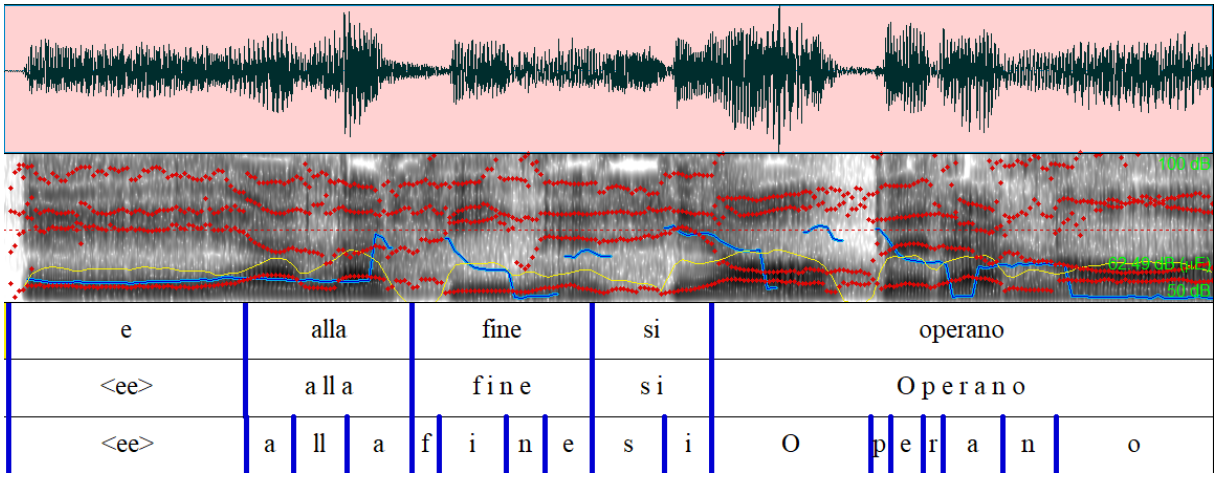


the decisions are statistically or deterministically taken, the linguistic and paralinguistic intent representations are generated. In the last module, tools are used to synthetize the voice with peculiar prosodic characteristics, according to the intent, correlating it to other paralinguistic aspects, such as gestures, facial expressions, and posture. In the next sessions, we will focus on available tools, which can be usefully exploited in the development of some of the above-described modules.

1.1 Input processor: Automatic Speech Recognition and Phonetic features

For the automatic speech recognition in the input processor module, the web service MAUS can be used to fulfil specific phonetic requirements. The Munich AUtomatic Segmentation (MAUS) system (Schiel 1999, Kisler et al. 2017) is a multilingual tool used to transcribe audio inputs and align transcription to the spectrogram. Beside the graphic transcription, it also provides the phonetical one in SAMPA for each word and each phone, as in Figure 2. In this framework, other meaningful phonetic features, which can be associated to the semantics of linguistic intents, can be extracted, such as intonation, pitch and intensity. Furthermore, starting from the processed output obtained from MAUS, sociolinguistic profiling of speakers can also be outlined by extracting information concerning the openness of vowels and other articulative peculiarities. The phonetic features to automatically extract from a spoken input to

Figure 2 - MAUS automatic annotation in Praat



enable the conversational to react to them or use them in a specific situations are annotated starting from the output obtained by such a service to be used for the training of the system itself.

1.2 Natural Language Understanding: Language Model

As far as the Natural Language Understanding module is concerned, different techniques can be used to develop language and semantic models, with which the machine can be provided. CLARIN supports the use of some linguistic tools for data collection and annotation. Among various techniques, the use of SRGS (Speech Recognition Grammar Specification)[2] is mostly preferred to assure the categorization of possible intents in a target-oriented dialogue system, with means of the description of each possible structure that can be uttered to express a particular concept. These grammars can be automatically extended, as far as lexical variability and inflectional morphology is concerned (Di Maro 2017), making use of semantic networks such as ItalWordNet (Roventini et al. 2000) and POS-tagging tools like TreeTagger (Schmid et al. 2007).

The language model to be used for conversational purposes can be enriched with pragmatic information. For this purpose, the Dialogue Act Mark-up Language (DiAML) could be used. Not only is it suitable to annotate the type of intent performed, but it is also useful in case we need to specify whether the user intent was merely dependent on the action motivating the dialogue itself, or whether it was a feedback to the previous turn (auto- and allo-feedback), or if it was signalling the turn-giving or turn-taking action, or opening, closing or structuring the conversation, or in case of social obligations adjacency pairs (Bunt et al. 2010). The specification of the performed act is indeed useful to improve the disambiguation and thus the understanding. For instance, knowing when a museum visitor is giving a feedback on something previously uttered by the guide or asking for more information or clarifications on the same concept is important to assure an appropriate reaction by the virtual agent. Other pragmatic phenomena can be manually annotated using tools as EXMARaLDA (Schmidt 2009), a system for the computer-assisted creation and analysis of spoken language corpora.

Data analysis can be both corpus-based and corpus-driven: on one hand a given corpus can help to confirm or refute a pre-existing theoretical construct (corpus-based), on the other hand a corpus can be used to generalise rules (corpus-driven). For modelling conversational interactions, spoken corpora are useful to capture all the pragmatic characteristics arising from dialogues. Therefore, a corpus-driven approach is preferably adopted. To achieve such aims, tools like SPOKES are truly interesting. SPOKES – currently available in Polish and English – is an online service for conversational corpus data search and exploration (Pezik et al. 2015). By exploring this corpus, information concerning the strategies used in conversation can be extracted to be modelled in a one's own language model. As a matter of fact, an Italian version of SPOKES is possibly desirable, starting from data collected through researches like the one presented in this paper. Providing pragmatic annotation in such tools is also an advisable goal to better be applied in the development of conversational agents. As far as the Italian language is concerned, an available dialogic corpus is CLIPS (Savy 2009), whose annotations can be used to extract linguistic and paralinguistic phenomena to be modelled.

1.3 Multimodal Fusion

The module responsible for the fusion of different channel of intents communication – spoken language and paralinguistic features, specifically gestures and prosodic profiles – can rely on data synchronised with a tool like ELAN, before being modelled through probabilistic rules. ELAN is a tool designed to annotate audio and video files (Wittenburg et al. 2006). In ELAN's tiers, TextGrids obtained with MAUS can be imported and overlapped to the other pragmatic and paralinguistic information modelled. The fusion of the annotations can be used to process both the understanding and the generation processes. This tool is being used for the CHROME project to specifically model the way the gatekeeper would communicate cultural contents. After having recorded authentic tour guides, video and audio files have been synchronized in ELAN, where expert annotators are going to mark linguistic and paralinguistic phenomena. Fusing different channels of communication together in the modelling phase will result in a virtual tourist guide able to communicate as naturally as human ones.

[2] Speech Recognition Grammar Specification Version 1.0: https://www.w3.org/TR/speech-grammar/

2 Conclusion

In this paper, a brief overview of CLARIN's tool to be applied in the development of multimodal conversational agents has been presented. This framework is intended to be deepened as a PhD research project, which is part of the Italian National Project CHROME (Cultural Heritage Resources Orienting Multimodal Experiences)[3]. The development of other conversational annotated data to be made available for similar researches is a desirable part of the presented research.

References

Bunt, H., Alexandersson, J., Carletta, J., Choe, J. W., Fang, A. C., Hasida, K., ... & Soria, C. (2010, May). Towards an ISO standard for dialogue act annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Di Maro, M., Valentino, M., Riccio, A., and Origlia, A. (2017). Graph Databases for Designing High-Performance Speech Recognition Grammars. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers.*

Kisler, T., Reichel U. D. and Schiel, F. (2017): Multilingual processing of speech via web services, *Computer Speech & Language*, Volume 45, September 2017 (pp. 326–347).

Pezik, P. (2015, August). Spokes-a search and exploration service for conversational corpus data. In *Selected Papers from the CLARIN 2014 Conference*, October 24-25, 2014, Soesterberg, The Netherlands (No. 116, pp. 99-109). Linköping University Electronic Press.

Roventini, A., Alonge, A., Calzolari, N., Magnini, B., and Bertagna, F. (2000, May). ItalWordNet: A Large Semantic Database for Italian. In *LREC.*

Savy, R., Cutugno. F. (2009). CLIPS. Diatopic, diamesic and diaphasic variations in spoken Italian, in in M. Mahlberg, V. González-Díaz, C. Smith (eds.), *On-line Proceedings of 5th Corpus Linguistics Conference*, (p 213).

Schiel, F. (1999). Automatic Phonetic Transcription of Non-Prompted Speech. In *Proc. of the ICPhS* (pp. 607-610).

Schmid, H., Baroni, M., Zanchetta, E., and Stein, A. (2007). The Enriched TreeTagger System. In *Intelligenza Artificiale IV-2*. (pp. 22-23).

Schmidt, T., and Wörner, K. (2009). EXMARaLDA–Creating, analysing and sharing spoken language corpora for pragmatic research. Pragmatics. *Quarterly Publication of the International Pragmatics Association* (IPrA), 19(4), 565-582.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1556-1559).

[3] CHROME website: http://www.chrome.unina.it/