Spoken Conversational Search for General Knowledge

Lina M. Rojas-Barahona, Pascal Bellec, Benoit Besset, Martinho Dos-Santos, Johannes Heinecke, Munshi Asadullah, Olivier Le-Blouch, Jean Y. Lancien, Géraldine Damnati, Emmanuel Mory and Frédéric Herledan
Orange Labs, 2 Avenue de Pierre Marzin, Lannion, France
{linamaria.rojasbarahona,pascal.bellec,benoit.besset,martinho.dossantos,
johannes.heinecke,munshi.asadullah,olivier.leblouch,jeanyves.lancien
geraldine.damnati,emmanuel.mory,frederic.herledan}@orange.com

Abstract

We present a spoken conversational question answering proof of concept that is able to answer questions about general knowledge from Wikidata[1]. The dialogue component does not only orchestrate various components but also solve coreferences and ellipsis.

1 Introduction

Conversational question answering is an open research problem. It studies the integration of *question answering* (QA) systems in a *dialogue system*(DS). Not long ago, each of these research subjects were studied separately; only very recently has studying the intersection between them gained increasing interest (Reddy et al., 2018; Choi et al., 2018).

We present a spoken conversational question answering system that is able to answer questions about general knowledge in French by calling two distinct QA systems. It solves coreference and ellipsis by modelling context. Furthermore, it is extensible, thus other components such as neural approaches for question-answering can be easily integrated. It is also possible to collect a dialogue corpus from its iterations.

In contrast to most conversational systems which support only speech, two input and output modalities are supported speech and text. Thus it is possible to let the user check the answers by either asking relevant Wikipedia excerpts or by navigating through the retrieved name entities or by exploring the answer details of the QA components: the confidence score as well as the set of explored triplets. Therefore, the user has the final word to consider the answer as correct or incorrect and to provide a reward, which can be used in the future for training reinforcement learning algorithms.

[1] https://www.wikidata.org

2 Architectural Description

The high-level architecture of the proposed system consists of a speech-processing front-end, an understanding component, a context manager, a generation component, and a synthesis component. The context manager provides contextualised mediation between the dialogue components and several question answering back-ends, which rely on data provided by Wikidata[1]. Interaction with a human user is achieved through a graphical user interface (GUI). Figure 1 depicts the components together with their interactions.

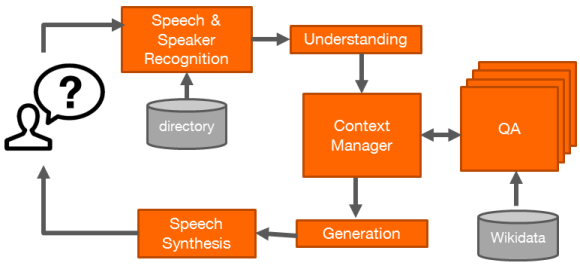

Figure 1: High-level depiction of the proposed spoken conversation question answering system. Arrows indicate data flow and direction.

In the remainder of this section, we explain the components of our system.

2.1 Speech and Speaker Recognition

The user vocally asks her question which is recorded through a microphone driven by the GUI. The audio chunks are then processed in parallel by a speech recognition component and a speaker recognition component.

Speech Recognition The Speech Recognition component enables the translation of

Proceedings of the SIGDial 2019 Conference, pages 110–113
Stockholm, Sweden, 11-13 September 2019.

speech into text. Cobalt Speech Recognition for French is a Kaldi-based speech-to-text decoder using a TDNN (Povey et al., 2016) acoustic model; trained on more than 2 000 hours of clean and noisy speech, a 1.7-million-word lexicon, and a 5-gram language model trained on 3 billion words.

Speaker Recognition The Speaker Recognition component answers the question "Who is speaking?". This component is based on deep neural network speaker embeddings called "x-vectors" (Snyder et al., 2018). Our team participated to the NIST SRE18 challenge (Sadjadi et al., 2019), reaching the 11th position among 48 participants.

Once identified, it is possible to access the information of the speaker by accessing a speaker database which includes attributes such as nationality. This is a key module for personalising the behaviour of the system, for instance, by supporting questions such as "Who is the president of the country where I was born?".

2.2 The Dialogue System

The transcribed utterance and the speaker information are passed to the dialogue system. This system contains an **understanding** component, a **context manager**, and a **generation** component (Figure 2).

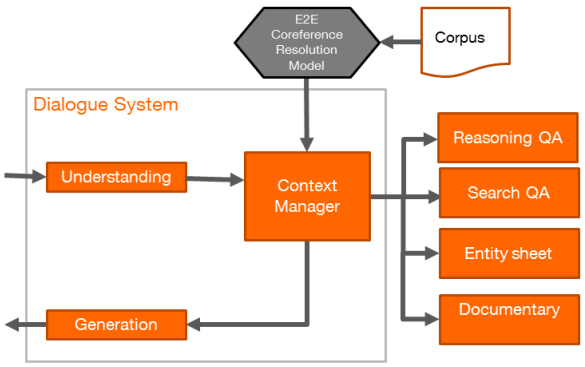

Figure 2: Internal structure of the proposed dialogue system, with emphasis placed on the interactions of the context manager.

Understanding The understanding component relies on a **linguistic module** to parser the user's inputs. The linguistic module supports part-of-the-speech (POS) tagging, lemmatisation, dependency syntax and semantics provided by an adapted version of UDpipe (Straka and Straková, 2017), extended with a French full-form lexicon. UDpipe was trained on the French GSD treebank version 2.3[2]. Since the syntax of questions in French differs from that of declaratives, we annotated manually about 500 questions to be merged into the UD treebank (which originally did not contain questions). Tests show that the labelled attached score (LAS) is thereby increased by 10% absolute, to 92%.

	Train	Dev	Test
words	208 245	45 001	89 330
sentences	10 166	2 976	4 853
mentions	15 013	3 008	6 232
incl. prons.	1 465	280	538
chains	3 793	901	1 533

Table 1: Subset of the corpus CALOR used for training, developing and testing of the coreference resolution module. Note that the values given for the mentions include pronouns.

Context Manager The Context Manager component is able to solve coreferences by using an adaptation of the end-to-end model presented in (Lee et al., 2017), that we trained for French by using fasttext multilingual character embeddings (Bojanowski et al., 2017). The data used to train the coreference resolution model is a subset of the corpus CALOR (Marzinotto et al., 2018) (Table 1), which has been manually annotated with coreferences. This corpus contains coreference chains of named entities, nouns and pronouns (such as "the president" – "JFK" – "he" – "his").

The dependency tree and semantic frames provided by the linguistic module are used to solve ellipsis by taking into account the syntactic and semantic structure of the previous question. Once the question has been resolved, it calls the QA systems and passes their results to the generation module.

Generation The generation component either returns the short answer provided by QA systems or relies on an external generation module that uses dependency grammar templates to generate more elaborated answers.

2.3 QA Systems

Two complementary question answering components were integrated into the system: the Reasoning QA and Search QA. Each of these

[2] http://universaldependencies.org/

QA systems computes a confidence score for every answer by using icsiboost (Favre et al., 2007), an Adaboost-based classifier trained on a corpus of around 21 000 questions. The Context Manager takes into account these scores to pick the higher-confidence of the two answers.

Besides the QA components, there are two other components that are able to provide complementary information about the Wikidata's entities under discussion: Documentary and Entity Sheet.

Reasoning QA The Reasoning QA system first parses the question by using a Prolog definite clause grammar (DCG), extended with word-embeddings to support variability in the vocabulary. Then it explores a graph containing logical patterns that are used to produce requests in SPARQL[3] that agree with the question.

Search QA The Search QA system uses an internal knowledge base, which finely indexes data using Elasticsearch. It is powered by Wikidata and enriched by Wikipedia, especially to calculate a Page-Rank (Page et al., 1997) on each entity. This QA system first determines the potential named entities in the question (i.e. subjects, predicates, and types of subjects). Second, it constructs a correlation matrix by looking for the triplets in Wikidata that link these entities. This matrix is filtered according the coverage of the question and the relevance of each entity in order to find the best answer.

Documentary The documentary component is able to extract pertinent excerpts of Wikipedia. It uses an internal documentary base, which indexes Wikipedia's paragraphs by incorporating the Wikidata entity's IDs into elasticsearch indexes. Thus, it is possible to find paragraphs (ranked by elasticsearch) illustrating the answer to the given question by taking into account the entities detected in the question and in the answer.

Entity Sheet The entity sheet component summarises an entity in Wikidata returning the description, the picture and the type of the entity.

[3] https://www.w3.org/TR/sparql11-query/

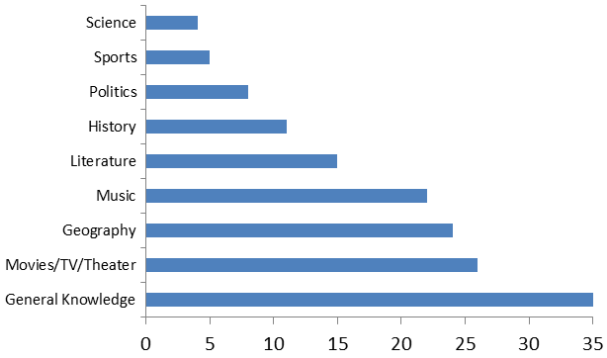

Figure 3: Distribution of question topics used to evaluate system performance on out-of-context questions.

2.4 Speech Synthesis

Finally, the generated response is passed to the GUI, which in turn passes it to the Voxygen synthesis solution.

3 Evaluation

The evaluation of the individual components of the proposed system was performed outside the scope of this work. We evaluated out-of-context questions, as well as the coreference resolution module.

Performance on out-of-context questions was evaluated on Bench'It, a dataset containing 150 open ended questions about general knowledge in French (Figure 3)[4]. The system reached a macro precision, recall and F-1 of 64.14%, 64.33% and 63.46% respectively[5].

We also evaluated the coreference resolution model on the test-set of CALOR (Table 1), obtaining an average precision, recall and F-1 of 65.59%, 48.86% and 55.77% respectively. The same model reached a average F-1 of 68.8% for English (Lee et al., 2017). Comparable measurements are not available for French. F-1 scores for French are believed to be lower because of the lower amount of annotated data.

4 Examples

On the one hand, the system is able to answer complex out-of-context questions such as "What are the capitals of the countries of the Iberian Peninsula?", by correctly answering the list of capitals: "Andorra la Vella, Gibraltar, Lisbon, Madrid".

[4] Publicly available in https://github.com/lmrojasb/benchit.git

[5] Following the metrics of the task-4 of QALD-7 https://project-hobbit.eu/challenges/qald2017/

U: Who is Michael Jackson ?

S: Michael Jackson is an American author,composer, singer and dancer

U: What is his father's name?

S: Joseph Jackson

U: and his mother's?

S: Katherine Jackson

U: and his brothers' and sisters'?

S: Tito Jackson, Rebbie Jackson, Randy Jackson, Jackie Jackson, Marlon Jackson, La Toya Jackson, Jermaine Jackson, Janet Jackson

Figure 4: English translation of French conversation involving in-context questions.

On the other hand, consider the dialogue presented in Figure 4, in which the user asks several related questions about Michael Jackson. First she asks "Who is Michael Jackson?" and the system correctly answers "Michael Jackson is an American author, composer, singer and dancer", note that this is the generated long answer.

The subsequent questions are related to the names of his family members. In order to correctly answer these questions, the resolution of coreferences is neccesary to solve the possessive pronouns, which in French agree in gender and number with the noun they introduce. In this specific example, while in English "his" is used in all the cases, in French it changes to: *son* père (father), *sa* mère (mother), *ses* frères (brothers). This example also illustrates resolution of elliptical questions in the context, by solving the question "and his mother's" as "What is the name of his mother".

5 Conclusion and Future Work

We have presented a spoken conversational question answering system, in French. The DS orchestrates different QA systems and returns the response with the higher confidence score. The system contains modules specifically designed for dealing with common spoken conversation phenomena such as coreference and ellipsis.

We will soon integrate a state-of-the art reading comprehension approach, support English language and improve the coreference resolution module. We are also interested in exploring policy learning, thus the system will be able to find the best criterion to chose the answer or to ask for clarification in the case of ambiguity and uncertainty.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of EMNLP 2018*, pages 2174–2184, Brussels, Belgium.

Benoit Favre, Dilek Hakkani-Tür, and Sebastien Cuendet. 2007. Icsiboost. `https://github.com/benob/icsiboost`.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 EMNLP*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Gabriel Marzinotto, Jeremy Auguste, Frédéric Béchet, Géraldine Damnati, and Alexis Nasr. 2018. Semantic frame parsing for information extraction: The CALOR corpus. In *LREC*, Miyazaki, Japan. ELRA.

Larry Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1997. Pagerank: Bringing order to the web.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahrmani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Proceedings of INTERSPEECH, 2016.*

Siva Reddy, Danqi Chen, and Christopher D Manning. 2018. Coqa: A conversational question answering challenge. *arXiv preprint arXiv:1808.07042.*

Seyed Omid Sadjadi, Craig S. Greenberg, Douglas A. Reynolds, Elliot Singer, Lisa P. Mason, , and Jaime Hernandez-Cordero. 2019. The 2018 nist speaker recognition evaluation. In *Proceedings of INTERSPEECH (submitted), Graz, Austria, August 2019.*

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khundanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *Proceedings of IEEE ICASSP, April 2018.*

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. ACL.