SotA - Chatbot for IR

Adrien Pouyet & Laure Soulier

1 Chatbots for Information Retrieval

1.1 From Information Retrieval to Conversational IR

As creatures gifted with consciousness and curiosity, and as workers in companies, we have information needs day to day. From cooking recipes to technical reports we seek new information continuously. To answer this human need with technologies, search engines have been developed since 1971[**?**]. Search engines are a great tool when you know how to express your need (eg : "ratatouille recipe") but they are not so great for complex queries (eg : if you want to know what classes you should take to complete a PhD in machine learning). Expressing an information need is often a complex task and can not directly be taught. Paradoxically, if you can perfectly express your information need it means you would probably not need the information. A simple illustration of such a paradox is when you want to discover a new domain. You usually do not have the right keywords and you will need several **searches − interactions ???** to find the words that will conduct you to the information you are seeking for. This is what we call section and it is illustrated in **?? cette phrase ne dit rien de particulier. La fusionner avec la suivante ?**. All technologies related to help users find an information or fulfil an information need falls under the name of Information Retrieval. Search engines such as Google, Yahoo, Bing, Yandex, Baidu, Naver or Qwant, are used everyday by billions of people but information retrieval is wider. First, a lot of websites have their own search engines. For example, on any e-commerce website the search bar hides a search engine. The technology may be very different from text engines because products have facets (ie : also known as tags) and prices that should be taken into account when a user is looking for a product. Another use case of information retrieval is image search. We actually use it day to day but it is important to notice that algorithms are very different when using text to search images compared to using text for text information ; it is even more different when you are searching images using images. For some years now, search engines such as Google propose to directly answer questions or present a summary of a Wikipedia page as the first result. This is a form of aggregation or summarizing which are trending topics in natural language processing and information retrieval communities. In certain aspect, question answering technologies also fall into information retrieval. Now that we have an idea of what information retrieval is, let's analyze what are the current lacks of these systems.

We previously introduced what information retrieval is. We also introduced one of the biggest challenge for the community, user revealment. It is difficult for us, as humans, to express correctly an information need. We all experienced going to some friend with an information need and had a hard time to find how to ask for it. And our friend usually had a hard time to actually understand what we wanted to know. Now, imagine doing the same task but having only one sentence for your search. Actually, it would not be a sentence as we express everyday but only keywords. This becomes a real challenge. Actually, community have shown[**?**] that 37% of queries are reformulation of previous ones[1]. From this observation, the community addressed the problem by proposing tools in search engines such as query suggestion or query expansion. Even though these techniques improved the search system and can be exploited by users it is not satisfying. First, the user have to know the search engine will expand its queries using a method like Rocchio to exploit it, this is what [**?**] called *System Revealment.* Search engines are complex machines with a lot of features we can use but users are rarely aware of these. Second, tools such as query suggestion certainly helps the user to do new search and it may help in discovering new keywords but it does not permit the user to detail its information need. *User Revealment* is the most critical aspect in information retrieval. If the user can perfectly reveal itself then the problem becomes way easier. Studies have shown that the fact of having a conversation is probably more important than what has been said by the helper (a librarian in the studies)**REF**[].

[1] This study is 10 years old, note that this percentage may have changed

From this last assumption, what is commonly known as *chatbots* may solve the problems we talked about. In the field we would rather speak of *Conversational Systems* and in the context of information retrieval we will talk about *Conversational Search Systems*. An important question to ask is whether chatbots are always useful or not. Literature, and we agree with it, determined that lot of information seeking tasks are better completed not using chatbots. For example booking an hostel is better done using a form on a website rather than filling slots when a system asks for it. In [**?**], they describe scenarios they identified as improvable using conversational search systems . They specify three scenarios : solving ambiguity, helping to specify important criteria and exploratory search. These are very generic tasks that can be further detailed and these scenarios are not exclusive either. Several works proposed complementary definitions for conversational search systems . We will go through those definitions from the most high level to the details. But first, let us define what a conversational system is.

1.2 Background and definitions

The definition of **conversational search systems** is inspired by [**?**, **?**, **?**], though we regrouped some propositions and chose a different structure. First, we will describe the properties we want the system to have. We define as a *property* a very high level feature, these are conceptual and not implementable features. Second, we list tasks the system has to do. We define tasks as as non visible processes the system has to do (e.g. : understanding natural language). Last, we will define actions the user and the system may take. We define actions as visible processes during the conversation (e.g. : asking a question but not how it is done).

1.2.1 Properties

User Revealment As previously explained, it is very difficult for users to express their needs. Without a proper need expression and user revealment, we can not fulfil the task. User revealment may take various forms. The simplest scenario is when the user may perfectly express its needs through a more or less complex query. In this case, a conversational search system is not useful. A more interesting scenario is when the user express a query that has partial information need and several queries will follow. Such an example is presented in Dialogue 1. Even though it is more complex, a conversational search system might not be that useful since it is already the behaviors of users in current search engines. Useful scenarios are more open. In dialogue 2, the initial query needs explanation on what the user is actually looking for. In order to help users to reveal themselves, **REF** analyzed what actions user did during a search conversation. Some of these actions can only be from user initiative but others may be requested by the system. For example, the conversational search system may ask for a critique over a proposed solution. We will detail these actions later. Another aspect of user revealment is showing users how to exploit the system.

USER: I would like to visit a dinosaur museum
USER: I'd like to visit a dinosaur museum in Paris

Dialogue 1: Sample of a query requiring other queries

USER: I am interested in the impact of the climate change on the GDP for both southern and northern countries

AGENT: I identified two search spaces that might interest you, either the positive changes (sustainable development, renewable energies, farming, ...) or the negative changes (farming, wild life, financial crisis, collapsology, ...). Which one are you interested in ?

USER: I already know about some of the positive impacts (I actually worked on sustainable developement and how to change farm methods for several years), I'd like to know about financial crisis. What is collapsology ?

AGENT: Collapsology is [...]. If you want to learn more about it, I suggest [...]

AGENT: I found some research works from University of [...]. Do you want me to find some examples of financial crisis provoked by climat change ?

USER: Thanks for the refs, I'll check that.

Dialogue 2: Example of a complex query and how an agent could react

System Revealment Search Engines, and more globally information retrieval systems, are very complex machines very mysterious to their users. If we take Google as an example, not all users know that using double quotes around several words forces the search to match the exact string. Even less users know we can exclude some words or expressions from the search. And Google has a lot more features I am not even aware of. All the features of web search engines are presented somewhere in a raw documentation no one wants to read. The less the users are comfortable with search engines, the more conversational search systems will be useful. So it is important to have a proper revealment from the system to help the users. Features of the conversational search system should be presented to users when it is useful for them to use. If the user is only doing text searches, it is not relevant to present a feature about image recognition. This is a big challenge since the algorithms should know what feature is relevant to better find an information without really knowing the information the user is looking for. This can be done after the search but it is not always an appropriate strategy. In order to have a good system revealment and to be more pleasant, it is important to have mixed initiative between the user and the system.

Mixed Initiative As we explained before, the fact of having a conversation is at least as important as the strategy the helper is using. A conversation is defined by the Cambridge Dictionnary **REF** as *a talk between two more people in which thoughts, feelings and ideas are expressed, questions are asked and answered, or news anf information is exchanged.* Even if the definition is about vocal discussion, the rise of text chats (MSN Messenger, Facebook Messenger, WhatsApp, Telegram, Discord, Texts, IRC, ...) leads to have the same definition for written discussions. So in the case of conversational search system , talking about thoughts, feelings or ideas is not appropriate. Actually [**?**] make the assumption we want to minimize the number of exchanges to help the user, exchanging thoughts does not seem appropriate for a search tool. So we focus here on the end of the definition *news and information is exchanged.* An exchange is two ways. So if having a conversation is important in the search process, we want the conversational search system to be pro-active. Mixed initiative may take several forms. As proposed by [**?**], the agent should complete side-tasks if it may help in the future. For example, if the user identify someone on a picture with several persons, agent may ask the user who are other persons. If we take the example of dialogue 2, the system may ask on what exactly the user worked on so the system may use this in the context of the future searches. Another type of mixed-initiative is when the system will ask the user an explicit critique of the answers retrieved (we will talk more about critiques later). Those examples introduce the concept of memory for conversational search systems .

Memory Having memory in a conversational system is crucial to make it interesting. If we talk with someone we want him/her to remember what we already said during the conversation. And if we have a friendship we expect our friend to remember what we experienced together. As researchers on conversational search systems , we have to take into account what makes a conversation entertaining and interesting. On top of that, from a search perspective, we already use search sessions to better understand the user's information need so we already know it is useful. Memory is about the current conversational search, so the agent knows what were the *Past Information Need* (PIN)[**?**] so it should try not to answer these needs but be more focused on the *Current Information Need* (CIN)[**?**]. It should also remember the past sessions since the user may have revealed important information to help the agent have a better understanding of the user's need. Another crucial aspect of memory in conversational search systems is to allow the user to correct what was said before. The user may have done a mistake and the conversational system should be able to change information in its memory.

Explainability Explainability and interpretability are trendy research topics. A lot of industries want to have proofs that algorithms will actually do what they are supposed to do and it may be complicated when having complex machinery. Conversational search systems are complex machinery and users may want to understand why and how the result has been found so they can be better at searching next time or just for the sake of curiosity. Depending on the technique used, and this thesis is focused on deep learning, it can be one of the hardest problem to solve for conversational search systems . Works on strong queries are an example of explainability. Finding the query that rank a document at the top is a useful feature for users and is a good example of how interpretability can be used to help the user in his search.

Personality and Moral Responsibility Giving a personality to conversational systems is both a complex search task and a questionable topic. Two of the papers we used for the definition [**?**, **?**] disagree with whether a conversational search system should have a personality and have moral responsibility.

Having a personality may help the user for the conversational aspect ; on one hand users could be more willing to reveal themselves if they think the chatbot as a human being. On another hand, is summons a lot of ethical questions that are not yet solved. More responsibility is a less controversial. The most used example is how should a system react when a user is looking for how to suicide. These are interesting aspects we will not address in this thesis.

General Knowledge Last but not least, general knowledge is an important property for conversational systems to make them interesting to interact with. Having common sense and general knowledge leads to have a context on the users' needs and make assumptions on what the user knows. If we know the user is older than 18 years old for example, we can assume that basic mathematics or history knowledge is known. From general knowledge we can also infer things that should not be said (e.g. : racist information, fake news, ...). This would also be a base for having a personality and moral responsibility. Again, we will not address this problem in this thesis.

1.2.2 High Level tasks

Now that we know what properties we want for our conversational search systems , we will define what tasks it should complete. In the following paragraphs we will define all high level tasks an agent should be able to do so the properties are complete. Note that high level tasks are research problems by themselves (so will be some of the lower level tasks) and will probably be addressed first by lower level tasks. Some of these asks, as well as the properties, are not specific to conversational search systems but are common to all conversational systems.

Natural Language The first issue users have with current chatbots is how natural language understanding and generation is handled. Having a good natural language understanding is now easier with current models **REF**[**?**] but we still lack good metrics to evaluate how good models are. Natural language understanding is the most important task a chatbot has to solve. Without a proper understanding, it can not understand what the user wants. In the same way, natural language generation is crucial to have interaction with users. If the generated sentences are not fluently readable for the user, it will be too frustrating and the conversational system will not be used. Handling natural language is not a new task in the natural language processing community. However, current metrics do not really measure how good the language is, we don't know how to. Given a text and an information need we can't measure if the need has been fulfilled. Natural language for information retrieval adds more challenges. The most important one is handling noise around the information need.

Filtering out superflous information Natural language is complex and not designed for computers. A lot of information conducted through natural language is implicit whereas a lot of words do not bring useful information or may be misleading. This leads to have a good management of filtering out superflous information for conversational systems . An example of query with superflous information and noisy natural language is presented in dialogue 3. Noisy queries are not restricted to natural language. It is an active research fields with several solutions **REFS + peut être expliquer**. Filtering information and handling natural language are kind of sub-tasks for intent identification. We chose to present those at the same level of identification intent because they are complex tasks by themselves and intent identification is wider than using natural language processing since queries may be multi-modal.

> USER: Hey, I'd like to eat in a french restaurant during a Paris trip. My mom never went to Paris before, I'm so excited to show her french food, especially snails ! I'd like to eat in the restaurant in a not so touristic place. I'll be in a hostel in Montmartre but I don't care on going anywhere in Paris.

Dialogue 3: Example of how natural language may be complex

Intent identification Once queried, the main objective of the chatbot is to identify what the user is willing, what the intent is. Intent may have very various forms. The user may be looking for a factual and scientific information, meaning it should be grounded with research papers or popularization articles. Or the user just wants an answer without proper grounding. Or the user is willing to discover a new area and the system should provide set of items rather than just one or two documents. We will detail identified intent later but this gives an idea of the kind of final answer the user wants. On top of that the conversational search system has to identify what the user is looking for, not only which form. This second aspect is currently more addressed in the literature **REF**. It requires a good natural language understanding and nice filtering when the input is in natural language . We are not interested in this thesis on other input than natural language but this intent identification should handle having images, audio or videos (not exhaustive) accompanied with text as query. Once the user intent is identified, the system has to answer.

Determine system response As explained in the previous paragraph, the user has some expectations on the final answer's form. Obviously, the conversational search system should identify it and tie to it. But during the conversation, the system has a lot of possibilities (more details in section 1.2.4) to answer and has to choose one. The first decision it has to make is whether the information provided is enough or if it needs a clarifying question. This is already a big challenge. Once this has been decided, the conversational search system has to choose the form of the answer : should it be ranked list or a single item ? Should it require more details on a precise point or an entire critique on the past answers ? This task is one of the most complex for the system. First, there is no limit on the type of task it should be able to do. A lot of works **REF** tried to elicit all possible actions but new works always identify new ones or slightly different ones. Second, for now it is not possible to evaluate if the decision was a good one. No dataset exists yet and producing one will be very costly. It would require to have several sample of each type of response (when it makes sense) per similar query and have enough variance in queries so the dataset is interesting. On top of choosing appropriate responses, the system has to manage the dialog so the user is still interested and the conversation is not frustrating.

Dialog management During any exchange there is always someone (or several persons) leading. Leading a conversation does not mean monopolizing the discussion but rather making sure everyone can speak and everyone is included in the conversation. In the conversational systems perspective, dialog management means the agent has to make sure the user is interested in the ongoing conversation. It implies the agent should give appropriate responses to the user with a proper tone. It should also adapt to the user's preferences. Dialog management is an active research field **REF** and is not restricted to conversational search systems but to all conversational systems .

Remember Choices In the wanted properties of a conversational search system , memory is a core one. During the conversation the user will reveal an information need and make choices. If we take holiday planning, the user will probably evoke a country and choose one or several cities to visit. These are crucial information the system need to remember. This is information for the current information need and within a single conversation session but the system also has to remember past information need of the session and any information useful for searches the user revealed. For example, if the user identified someone on a picture the system should never has to ask again for this information in a different session.

Document aggregation Last high level task we identified is aggregating documents. Document aggregation encapsulates a lot of outcomes. Aggregation might be a summary**REF** of several documents we present to the user for example. It could be useful when the system identify two area in the result space, summarizing the sub-spaces and presenting these to the user if a form of clarifying question. Another aggregation is presenting documents that do not have the same form (e.g. : text, music and pictures) but that complete the user need. An example widely used **REF** [**?**, **?**] is holiday planning. Holiday planning requires finding a destination (a country, a city, a region, ...), a travel, one or several places to sleep, places to visit, some pictures of the places, ... Retrieving such a set of document is a very complex, especially in open domain, that needs a lot of investigation to be relevant in industry application.

1.2.3 User Actions

Now that we know what are the properties and high level tasks of a conversational search system , we will focus on what actions may happen during the conversation. We first look in details at the users' actions before explaining system's actions. Note that this section is inspired by [**?**] though we did not choose the same aggregation of actions[2].

[2] We do not motivate here why we chose a different structure.

Reveal Actions The main task of the user when using any search system is to reveal its information need. Revealing goes through *disclosures* with various forms. We give some examples of disclosures in dialogue 4. Disclosure is not the only form of revealment from the user. Once information was given by the user, it can still be changed through *reexamination*[3] or *refinement*. Changing a statement is fairly usual in human conversations. If we take a task such as holiday planning, users may finally increase their budget or add a traveller in the planning. Finally, user may *expand* its information need by widening its interest like adding a city to the search possibilities. Reveal actions also encapsulate all *critiques* the user may do. Critiques may have various forms[**?**] such as free text, note, ...

> USER: I would like to discover how GMOs are made. *[Disclosure - Volunteer]*
> USER: Ok, but I'm not interested in techniques like CRISPR-Cas 9. *[Disclosure - Not]*
> USER: I am not sure if I want to visit a lab working on GMOs, it seems out of reach to me. *[Disclosure - Unsure]*

Dialogue 4: Disclosures example. We can think of other types of disclosures, these are not exhaustive.

Request Actions Once information was disclosed, the system will probably present some results. With these results the user may ask for a detailed *set* or *ranking* of items. The user may also want to have a *summary* of several items. For example if the system says there is two subspace in the search results, user may want to ask for a summarizing so it is easier to choose. Given several results, it is likely users will ask for *comparison*, *subset* selection, *detail* options or asking for *similar* results. Note that a lot of these actions should not be required in an ideal world, we expect from the system to know what kind of results the user wants. However, some actions such as *repeat* to revisit an option or *back* to previous option can not be guessed by the system but are still useful from user perspective. Finally, users may want to have some *explaination* about why an item was proposed or ask the system what it *understands* about the current (or past and future) information need.

1.2.4 System Actions

Users have actions they intuitively do. Previous works **REF** worked on identifying these actions so it is easier for us, researchers, to know what to expect and it provides a more formal framework. But agents' actions are to be defined and can not be expanded intuitively. Users can find new ways or new requests slightly different than the one we identified but for now systems can't. Thus we have to define precisely what a system can and can not do with the objective of completing the properties and high level tasks we defined above as well as making sure the user can do any kind of action. In the following we detail actions we identified from the existing literature. A lot of the system actions are the counterpart of user actions.

Reveal Actions Users need to reveal their needs and systems need to reveal both their results and their features. Presenting the results is another research field (human-computer interaction) but here we consider which form the result might take : *set* of item, *ranking* of documents, *summarizing* documents, *partial* items or *choices* between several options (this is not exhaustive, though we tried to be). A lot of these actions were motivated before but let us bring back these in the context of actions. We illustrated them and the following ones in dialogue 5. Showing a set of items (without an order) may be used for suggestion purpose or for task the requires it (e.g. : result of a holiday planning). It can also be used if the ideal document does not exist and several items are necessary to fulfill the user's need. Presenting a ranking is widely used nowadays. It is still relevant in the conversational search systems framework since a conversational search system is not perfect in understanding the information need nor it is to retrieve documents. We see summarizing documents mainly used to present sub-spaces of the results. A good example was presented in dialogue 2 where the system identifies either positive or negative impacts. Presenting partial items is appropriate when only part of the facets of an item is relevant for the information need (e.g. : size of a television but not the brand). On top of presenting results, system may *demonstrate* features when appropriate. The system can answer to an explainability request by revealing what the system *understands* of the information need or *explains* why a an action was taken (or an item chosen). Finally, a conversational search system should identify its weaknesses and be able to express its *doubts* and acknowledge the *errors* to the user so the user better understands how it works.

[3] In [**?**] they use the term *revise*

User Revealment	System Revealment	Mixed Initiative	Memory	Explainability	Personality/ Morality	General Knowledge
disclosure reexamination refinement **critique** expand	set ranking summary **understand** **explain** partial **demonstrate** **doubts** error	**demonstrate** doubts errors **clarify** **critique** **suggest** **long-term**	repeat back explain **understand**	**explains** doubts error **understand**	Non-relevant	Non-relevant

Table 1: Overview of what action contributes to which properties. Is is not relevant for personality/morality and general knowledge to list actions since it can not be shown through specific actions but rather is a general point of view.

SYSTEM: So for your holidays, I identified you want to go to France by train from 10th February to 22th February and be in Paris for Valentine's Day. I planned a train from Lisbonna to Paris at 6am on 10th Feb. and from Paris to Madrid at 10am on 22th Feb. The hostels Jardin du Marais is for the Valentine's Day and [...] *[Set of items]*

SYSTEM: The search results of ”dinosaur museum in Toulouse region” retrieved the following results : [...] *[Ranking]*

SYSTEM: I can identify two subspaces in the search area. One is talking about what people fear about nuclear powerplants, the other is more technical and explains how power is produced. *[Summarizing documents]*

SYSTEM: I found two references you might like. One has a diagonal of 55 inches and is in 4k, the other is 65 inches and is OLED. Which one you prefer ? *[Partial items and Choices]*

SYSTEM: You can exclude words from my searches explicitly like : ”restaurants NOT [italian indian]” *[Demonstrate]*

SYSTEM: If I understood correctly, you want to learn about second degree polynomial solving algorithm but you don't know complex number nor how factorisation works. *[Understands]*

SYSTEM: I retrieved the article ”Attention Is All You Need” because it is widely cited in the natural language processing community. *[Explains]*

SYSTEM: I think you might like eating in a Bouillon like the one in Pigale. *[Doubts]*

SYSTEM: I failed to understand your need, I think I am not qualified to answer medical questions. *[Error]*

Dialogue 5: System Revealment

Request Actions Since the agent should be aware of what it understands from the information need and whether it is clear enough to do a search or not, a conversational search system can ask *clarifying* questions. These questions may have several form such as repeating what the user said in a different form or disambiguate a term. We described in the properties a conversational search system should adapt to the user and improve over times, then it is important it can ask for *critiques* as feedback. Finally, we want the agent to be pro-active in the conversation. One of the possible way is to *suggest* sub-spaces of the search area whereas it is similar to what the user is looking for (e.g. : suggest Spain if the user wants to go to Italy) or after an abstraction of the current information need, the system may suggest some slots that may need to be filled the user did not evoked yet. Finally, the system may deviate from the conversation and ask for *long-term* information such as identifying if the user is vegetarian or not during a restaurant search.

Natural Language	Filtering	Intent Identification	Determine System Response	Dialog Management	Remember Choices	Document Aggergation
disclosures reexamination refinement expand **critique** **summary** **explain** **understand** doubts clarify suggest	**disclosures** reexamination refinement expand **critique** detail	**disclosures** reexamination refinement expand **critique** understand explain	set ranking summary comparison **partial** **subset** similar choice **demonstrate** doubts error **clarify** critique suggest **long-term**	**critique** summary set ranking partial choice **demonstrate** **doubts** **clarify** **suggest** **long-term**	**set** ranking summary partial choice demonstrate **clarify** **understands** **explains** suggest long-term	**set** ranking **summary** **partial** choice **explains** critique **suggest** comparison **subset** similar

Table 2: Overview of what action contributes to which task. Note that not all actions from users are listed here since it might be irrelevant (like *repeat* or *back*)

- Comment est-ce défini dans le sota ?
  - Les tâches pour lesquelles un conversational search system c'est pas pertinent

TODO

- Ajouter des références
- remplacer les fausses références par un fichier .bib

1.3 Challenges / Etat art

1.4 Evaluation for Conversational Search Systems

We can design the most sophisticated models, if you can not evaluate and compare them, it is pointless research. Having a good evaluation protocol is crucial in every research domain and we will see that for conversational search systems it is not obvious how to properly evaluate at minimal cost. Conversational Systems can only be evaluated through human data since it is designed to be used by humans. Currently, we can not simulate users with satisfying results so we have to design models by considering we will not have access to human feedback during the learning process. It is possible to have human evaluation at the end but it is expensive so it usually involves few people. In the literature, there are three types of datasets for conversational system : machine-machine dialogues, human-machine dialogues and human-human dialogs. As detailed by [**?**] 1) machine to machine dialogues are usually templates that are not realistic but have the advantage of being automatically generated (which leads to big datasets), 2) human to machine is not easily collected since we need to have interaction with a machine, it is usually used to improve systems rather to build and 3) human to human dialogues is the best for conversation since it is the objective of conversational systems to be human like. We will first analyse what are the metrics and challenging in evaluating conversational systems , then we will detail every existing dataset or environment for conversational systems and finally present all datasets available for conversational search systems .

1.4.1 Challenges & Metrics

exemples de tâches
Conversation response ranking conversation up to utterance x, a small set (10/50) of potential responses including the correct one; ranking of responses - Dialogue act prediction dialogue utterance; utterance label (e.g. further details) - Sub-goals prediction, clarification question generation dialogue utterance (extra-)topical aspects/facets - Sequential QA series of interconnected questions series of answer rankings

WIP Evaluation : goals are defined as slots that need to be filled. Usually there are several metrics used to evaluate a model in dialogs. We want to both measure how well the task was completed and whether the language was good or not.We measure the match of entities and the success rate of tasks.

WIP Wizard of Oz (and free texts. The idea behind Wizard of Oz datasets is to collect data from humans using a finite set of response. This finite set of responses can be template filling or retrieved from a database. To cite **??** : ”*The goal of a Wizard of Oz (WoZ) methodology is to mimic a hypothetical system which is not yet in existence or one which is bounded by technical limitations [...].*” The idea is to propose as many responses as possible in as many context as possible. Then, we can use Woz framework refs : kelley 1984 & Wen et al 2017 and Asri et al 2017

ref

- Natural Language Understanding : no good metric and what are the difficulties
- Human Interaction is needed for dialogs
- We get around this issue using slot detection

1.4.2 Conversational Systems

1.4.3 Conversational Search Systems

Now that we presented every conversational datasets non related to information retrieval, let us introduce the remaining ones.

Spoken Conversational Search pas sûr d'avoir la bonne ref

MISC This dataset is proposed by [**?**] and is designed for spoken conversational search. Pairs of participants were formed to complete a task. One of the participant was the seeker with an information need but did not have access to any resources. The other participant was an intermediary and had access to a web browser. The dataset is very small (88 tasks completed) but is accompanied with an analyse over emotions and how the participants felt during the task completion. Released data consists of both speechs, videos and text as well as what was done on the computer. It is not usable data for deep learning but it is an interesting paper to understand how users of conversational search systems may behave.

Frames Frames is proposed by [**?**] collected using the Wizard of Oz framework and is designed to study the role of memory in goal oriented conversational systems. This work is motivated by generalising the state tracking task. In state tracking, all information summarized a dialogue into a single ”frame” (a representation, usually slots) whereas they propose to have several frames to track choices and preferences (they take as example an exploratory search on e-commerce websites). The dataset is composed of 1369 dialogues with 15 turns in average. They also propose three tasks associated with the dataset : Natural Language Generation, Dialogue Management and Frame Tracking. Frame Tracking is having several frames to embed identified choices of the user during the conversation. Collected dialogues have turns where the wizard (the system) refers to previous frames. On top of that, it happens that the wizard executes several actions during an utterance, which is challenging for reinforcement learning algorithms. The Natural Language Generation task is illustrated by a *”The cheapest available flight is 1947.14USD”*. It means an agent should be able to generate sentences from a database query that embeds complex informations (”the cheapest”).

KVRET Proposed by [**?**], this dataset is designed for in-car personal assistant and is thus multi-domain. It consists of 3031 dialogs with 5.25 turns average grounded with a knowledge base. Data collections was performed using the Wizard of Oz framework. The objective here is to sustain a grounded discourse using a knowledge base and complete the tasks asked by the users. This dataset is a bit small but they proposed a deep learning network to fulfill the task.

éclaircir ce point, c’est pas clair la tâche du dataset

CoQA CoQA [**?**] is a Conversational Question-Answering dataset. It has been proposed to build models that are able to understand cross-references (such as : *”what is he running for ?”*) and is composed of 8399 dialogues with 15.2 turns in average and over 127k question-answer pairs. In this dataset a passage is given to both a questioner and an answerer (humans) and an historic of conversation. They have to ask questions and answer them. The answerer has to span the text from the passage that helped to answer. They also collected several responses for each question so conversations so following conversations would be different and the dataset richer. Finally, they propose an in-domain and out-domain evaluation.

MultiWOZ MultiWOz is proposed by [**?**] and is based on the Wizard of Oz framework we describe earlier. It is designed to be on multiple domain with conversations happenning on several domains (up to 5). It contains 10k dialogues (split into train-test-val) on seven different domains. For single domain conversation dialogues averages 8.93 turns and 15.39 turns for multi-domain conversations. The paper propose three tasks on the data : dialogue state tracking, dialogue context to text generation and dialog act to text generation. The firs task evaluate natural language understanding. Second task handles dialog management and response generation. Last task addresses generating natural language from structured meaning representation. Those tasks are quite representative of what a conversational search system has to do.

QuAC QuAC [**?**] is a dataset for Conversational Question Answering. A Wikipedia title is given to a student that has to ask questions to the teacher. Teacher has to answer by selecting text from the wikipedia page (optionnaly adding yes/no). It is composed of 13594 dialogues with an average of 7.2 questions per dialogue. Teacher may add an encouragement about continuing a line of questioning and was able to respond with a ”No Answer” token. This dataset contains cross-references so the agent solving the task has to use previous questions and answers to respond a question.

is it 7.2 turns or 15.4 ?, i’ll suppose 15.4

Wizard of Wikipedia Wizard of Wikipedia has been proposed by [**?**] and is based on the Wizard of Oz framework. The apprentice is willing to learn about the proposed subject (a Wikipedia webpage) and thus the use of knowledge from the wizard is encouraged (wizard is not an expert of the subject). Wizard’s utterances are grounded since they have to choose a proposed sentence to answer the apprentice. There are 22k dialogues with an average of 9 turns based on 1365 topics (each linked to a Wikipedia article) in open domain (Gouda Cheese, music festivals, ...). This makes the dataset having several dialogues for each topics wich is an interesting property.

ShARC ShARC [**?**] is a conversational question answering dataset. Contrary to previous conversational question-answering dataset we presented before this one uses ”ruled-based” documents. Rule based documents express, using natural language, how something works. They take as example legislation document to determine if you have to pay taxes or not. This kind of document is difficult because an

Name	#Dialogues	#Turns (avg.)	information need	multi-turn	multi-intent	clf. questions	inf. needs	utterance labels	multi-domain	grounded
SCS	39		●	●	●	●	●	●	●	○
MISC	88		●	●	●	●	●	○	●	○
Frames	1,369	15	●	●	●	●	○	●	○	○
KVRET	3,031	5.25	●	●	●	○	○	●	○	●
CoQA	8,000	15.2	●	●	○	○	○	○	●	●
MultiWOZ	10,438	∼12	●	●	●	●	○	●	○	○
QuAC	13,594	15.4	●	●	○	○	○	●	●	●
Wizard of Wikipedia	22,311	9	●	●	●	○	○	○	●	●
ShARC	N/A	32,436	●	●	○	○	○	○	○	●
MSDialog	35,000	4.56	●	●	●	●	●	●	○	○
DSTC 7	100,000		●	●	●	●	●	○	○	○
Ubuntu Dialogue Corpus	930,000	7.71	●	●	●	●	●	○	○	○

Table 3: Datasetsfor Conversational (Search) Systems and their properties.

agent has to deeply understand the text to be able to ask useful questions as well as giving the right answer. For the 32436 utterances they use only 948 documents. This is an interesting property since we have several dialogues from the same document. They organised dialogues within a tree with the possibility of "skipping" a node depending the question which is realistic. In the paper they detail metrics they used which makes the dataset suitable for benchmarking models.

MSDialog MSDialog [?] is a dataset built upon a Microsoft forum with over 35000 dialogues. Part of the dataset (2400 dialogues) is annotated on the user intent (such as *information request* or *clarifying question*) and the rest is left unsupervised. Dialogues contains multiple participants which is not usual in datasets we presented before. Proposed task is to identify intent of the participants but it can be extended to other tasks.

DSTC 7

Ubuntu Dialogue Corpus UDC [?] is the biggest available dataset for dialogues. It contains 930k dialogues with 7.71 turns in average wich makes it really suitable for deep learning. It only contains 2 persons discussion and is extracted from an online chat (IRC) about Ubuntu support. Even though it is not grounded for each utterance, using the documentation from Ubuntu and its packages seems possible to extend the task. The task is Best Response Selection, but again it could be extended through reinforcement learning or supervised learning to generate the answer.
TREC : Complex Answer Retrieval Ubuntu others ?

add refs, MANTIS ref too

add conversations only datasets

2 Ideas

- Sur la tâche de Strong Query (générer une requête qui fait remonter le document le plus haut possible) : on peut s'en servir pour s'évaluer sur notre traduction NL ¡-¿ KW : à priori avec des mots clés on peut faire mieux qu'en NL sur un BM25
- Modéliser le dialogue par un MDP : on passe d'une "original question" à "follow up question" par exemple -¿ Peut aider à prédire la réponse utilisateur et donc avoir une meilleure réponse