Building Coherent Open-domain Dialogue Systems: Data Collection and Analysis

Songbo Hu

4th Year Project Report
Artificial Intelligence and Computer Science
School of Informatics
University of Edinburgh

2020

Abstract

Recent advances in neural network based language models enable us to learn entire dialogue models directly from conversational data. However, most of the previous systems have failed to model the longer prior context of the conversation adequately, thereby creating a situation where extended dialogues rapidly become incoherent and unnatural sounding. There is no existing dialogue corpus which contains natural conversations and can provide strong supervision signals to train dialogue models with reinforcement learning. In this project, we design and collect a novel corpus contains extended coherent dialogues of questions and responses. These dialogues describe the relations within a knowledge graph with clear links between the phrases in the questions to nodes in a knowledge graph. The purpose of this corpus is to perform statistical analysis about various linguistic patterns within coherent dialogues, thereby creating a framework for neural models to learn how to parse these patterns and produce more coherent responses. We mainly focus on investigating when the appearance of an elided construction is natural and convey the intended content. A detailed statistical analysis has been conducted to support our investigation.

Acknowledgements

This thesis is the result of my undergraduate journey. Many people have influenced, inspired and supported me during the journey. First and foremost, I would like to thank my supervisor Prof. Alex Lascarides for providing guidance, feedback and constant help for this project. Her mentorship has not just greatly influenced my work on this thesis but will influence my research career in the future. I am also enormously grateful to all the staffs at the University of Edinburgh, my personal tutor Prof. Jane Hillston, all my lecturers, ITO, Advice Place... They offer me invaluable help and make me stronger during my journey. Special thanks to Informatics Student Services, who has funded this project and make the creation of our corpus possible. I appreciate all the help from whom I share this journey with, my lovely participants of this project, my friends and my colleagues at Edinburgh. Last, but certainly not least, I would like to thank my family in China for their selfless support and encouragement, without whom, my journey would never start.

Table of Contents

Chapter 1

Introduction

When was Alan Turing born?
Alan Turing was born on Sunday 13 June 1912.
Where was he born?
Alan Turing was born in Warrington Lodge.
Who is his father?
Julius Mathison Turing is Alan Turing's father.
When was he born?
Julius Mathison Turing was born 9 November 1873.
Where?
You're at Edinburgh, Scotland.

Songbo and Siri 2020

A dialogue system is a conversational agent that can converse with humans in natural language. It can help us solve many tedious tasks effectively or bring entertainment value to our daily life. Recent advances in dialogue systems (such as virtual assistants [2, 54, 5]) enable us to engage in natural conversational interactions and access to machine readable knowledge and information with computers. It substantially increases the accessibility of state of the art technologies to people with diverse backgrounds. And for those with impairments, it could transform their daily living into a journey toward capability instead of disability.

Since Alan Turing published his landmark work in 1950 [77], the intelligence level of a machine is described as how well the machine is able to fool a human into believing that it, the machine, is a human based on its text responses. If a human is unable to distinguish the difference between the machine from a human, the machine is said to have passed the Turing test. We could say such a machine signifies a significantly high level of intelligence. It has been the goal for the development of dialogue systems for decades, and various attempts have been proposed to pass the Turing test. Recently, advances in neural network based language models enable us to learn entire dialogue models directly from conversational data. Although we improve the performance of neural models by increasing the number of parameters and training with more data, we are still far from our goal. Most of the previous systems have failed to model the

longer prior context of the conversation adequately, thereby creating a situation where extended dialogues rapidly become incoherent and unnatural sounding, with repetition being a key problem.

(1) a. Q: When was Alan Turing born?
b. A: Alan Turing was born on Sunday 13 June 1912.
c. Q: Where was he born?
d. A: Alan Turing was born in Warrington Lodge.
e. Q: Who is his father?
f. A: Julius Mathison Turing is Alan Turing's father.
g. Q: When was he born?
h. A: Julius Mathison Turing was born 9 November 1873.
i. Q: Where?
j. A: You're at Edinburgh, Scotland.

As we demonstrated in the example conversation at the beginning of the chapter and repeated here, humans tend to utilise linguistic patterns, such as anaphora and Sluicing, to reduce repetitions. For example, in the human utterance (1c), human uses the pronoun "he" to refer "Alan Turing" in question (1b). This is a typical example use of anaphora, which leads to a more concise sentence. In addition, the human utterance (1i) is an example of sluicing constructions. Information, which can be inferred from the context, is omitted. A more refined definition of these linguistic patterns is provided in our Annotation Schema in Appendix A.

In contrast, dialogue systems rarely produce pronouns and unable to model elliptical constructions; therefore, produce unnatural speech in the context. For example, in order to answer the second question (1c), humans can interpret the anaphora (He refers to Alan Turing.) and produce a short answer (In Warrington Lodge.). Compared to human's responses, the Siri's response (1d) is heavy and verbose. The challenge for Siri, and many other dialogue systems, is to know when such an answer conveys the intended content, and when it does not. The system has to make a decent decision on when content can be elided while still ensuring that the user can easily interpret the system's message and retrieve that message accurately. Also, a wrong interpretation of a question could lead to an unhelpful and diverged response. Siri fails to understand the context to parse fragments and interpret the sluicing in the final question (1i). The misunderstanding of the context leads to a completely unhelpful and incoherent response (1j). The main reason for this unnaturalness for Siri is the lack of reliable systematic analysis of such linguistic patterns and the inadequate representation of the context. It is a problem which needs to be adequately addressed before our agents could look forward to attempting the Turing test.

We aim to address the gap in our understanding of the interaction between topic shifts on the one hand and the natural use of anaphora and elided constructions on the other by designing and implementing a method of data collection of humans engaged in question-answering in extended dialogue, where we control the way the topic shifts in

the course of a conversation. The resulting data provides an opportunity to systematically analyse the effects of topic shifts on the decisions that competent human language users make on whether to use anaphora and/or elided constructions. We motivate and describe the way we collected data in Chapter 3 and present an analysis of the data in Chapter 4.

In this section, we will briefly review the recent development of dialogue systems. Specifically, we will discuss two types of dialogue systems: the chit-chat dialogue systems (known as chatbots), and the task-oriented dialogue systems. We will focus on machine learning approaches to developing such systems, rather than the use of hand-crafted rules since the former methods are now dominant and the latter are highly brittle and restricted to toy domains. We will discuss their architectures, pros and cons, and their main limitations to achieve coherent open-domain conversations. The majority of this thesis is about how to create an environment for neural models to learn linguistic patterns within coherent extended dialogues and how to integrate these models into the existing dialogue systems. It constitutes a first step towards modelling coherent open-domain conversations and will potentially enable the existing dialogue systems utilising valuable linguistic information to produce coherent responses.

1.1 A Brief Review of Existing Dialogue Systems

1.1.1 The Chit-chat Dialogue Systems

The chit-chat dialogue systems are usually called chatbots. These systems are designed and implemented to carry extended conversations with the goal of mimicking the "chats" characteristic of informal human-human interaction [40]. They are mainly aiming to provide entertainment value. Existing chatbots majorly fall into the following subcategories: the rule-based systems, the corpus-based systems, as will be discussed in order below.

The Rule-based Systems

Instead of letting machines learn the conversation strategy from human behaviours, we could encode directly as rules. Dialogue agents could use these rules to generate dialogue utterances effectively. Normally, a message input will be processed by a set of carefully predefined rules, e.g. a key-word look-up dictionary, if-else conditions, or more sophisticated machine learning classifiers [44]. Consequently, the dialogue agent would produce a natural language response by outputting an utterance in storage, manipulating the input message or selecting some related historical contexts based on these rules.

There are many well-known rule-based dialogue systems in history, such as ELIZA [80] [1], SHRDLU [83], ALICE [79]. There are also languages such as Artificial Intel-

[1] It is instructive that the purpose of building Eliza was to show that it is sometimes possible to fake understanding via hand-crafted rules based on keyword search and template generation, even though there is no interpretation component or parsing in the system at all.

```
function ELIZA GENERATOR(user sentence) returns response

Find the word w in sentence that has the highest keyword rank
 if w exists
     Choose the highest ranked rule r for w that matches sentence
     response ← Apply the transform in r to sentence
     if w = 'my'
         future ← Apply a transformation from the 'memory' rule list to sentence
         Push future onto memory stack
 else (no keyword applies)
  either
     response ← Apply the transform for the NONE keyword to sentence
  or
     response ← Pop the top response from the memory stack
 return(response)
```

Figure 1.1: A Simplified Sketch of the ELIZA Algorithm.

ligence Markup Language(AIML), which provides a useful tool to write sophisticated conversations logic in a machine-readable format [79].

The development of rule-based systems is an important milestone in developing modern dialogue systems. Many of them are very sophisticated and well-engineered, which they attempted to model dialogue in an explanatory and deep way. The architecture of such a system involves the whole pipeline of natural language processing. However, their drawbacks are apparent: rule-based systems predominantly rely on the set of predefined rules, and these rules have to be carefully designed and implemented. Building a sophisticated rule-based system is very expensive because the number of these rules escalates rapidly. Rule-based systems do not have the ability to understand human languages and generate meaningful natural language utterances [44]. Consequently, they are very brittle and only able to conduct very superficial conversations. Despite the lack of intelligence, even with the recent rapid development of large fancy neural network architectures and the increasing number of conversational corpora, these rule-based dialogue systems could always provide a solid baseline.

The Corpus-based Systems

Coding conversation logics manually is astronomically expensive and infeasible for many applications. Corpus-based dialogue systems could potentially alleviate this issue by mining human-human conversations, or sometimes mining the human responses from human-machine conversations, instead of using hand-built rules. These data-driven approaches are becoming widely popular due to the increasing computing power and the creation of large scale conversational corpora.

There are two common architectures for such a system: information retrieval (IR), and machine-learned sequence transduction. Due to the limitation that IR-based chatbots can only mirror training data, it is often to treat response generation as a machine translation task which transduces from the user's prior turn to the system's turn. This

method offers the promise of scalability and language-independence, together with the capacity to capture contextual dependencies in a way not possible with IR-based approaches.

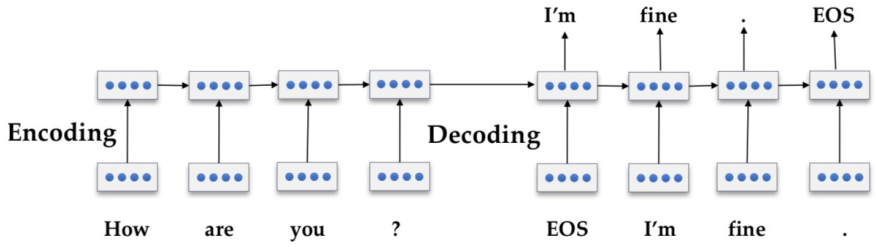


Figure 1.2: A Sequence to Sequence Model for Neural Response Generation [40].

This idea was firstly developed by Ritter et al. [67]. (cited in Jurafsky and Martin [40]) using phrase-based machine translation to translate a user turn to a system response. In 2015, transduction models for response generation were modelled using encoder-decoder (seq2seq) models [73, 76, 75]. However, the simple seq2seq generation architecture is unable to model the prior context of the conversation. Serban et al. [70] suggest a hierarchical model (HRED) that summarizes information over multiple prior turns. This model consists of two recurrent neural networks (RNNs) stacked on top of each other: one is a sentence-level RNN which encodes each utterance into a fixed-length vector, while a conversation-level RNN takes each utterance vector as input and outputs a vector that summarizes the conversation so far. The vector is mapped back to text using a recurrent decoder. This gives way for the previous information to be passed to future turns as hidden states [52].

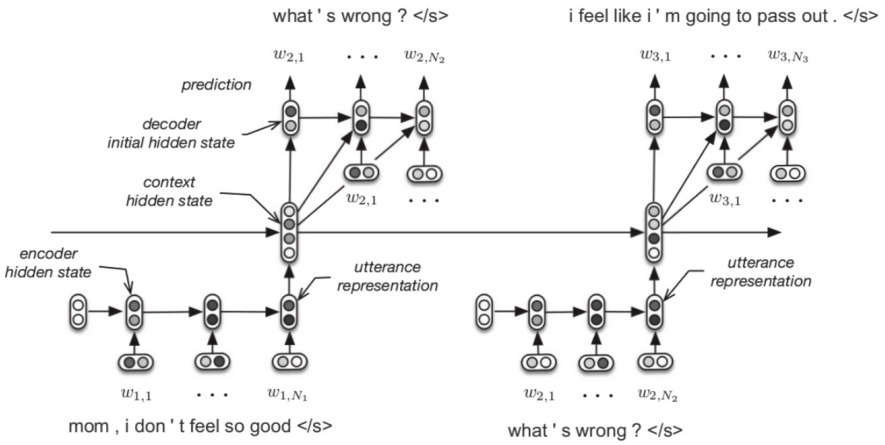


Figure 1.3: The Computational Graph of the HRED Architecture for a Dialogue Composed of Three Turns.

However, even if one has access to an enormous dataset for training, there is still a significant proportion of unseen dialogue states. Techniques, such as smoothing, can only help to a limited extent, because of the radical extent to which data is sparse. Consequently, generating responses using end-to-end methods tends to generate generic, uninformative and non-coherent replies (e.g., generating "I don't know." regardless of

the context). In addition, encoder-decoder response generators focus on generating single responses, instead of forming a coherent, continuous conversation [40]. Humans could quickly notice these unnatural and mechanical responses (e.g. the Siri conversation above). Techniques, such as reinforcement learning (RL) [45] and adversarial networks [46], can be used to address this issue.

1.1.2 The Goal-oriented Dialogue Systems

Goal-oriented dialogue systems are sometimes also called task-based dialogue systems, in which they converse with users to help complete tasks like making a restaurant reservation, booking a hotel or setting up an alarm. Famous examples of such agents are digital assistants (Siri, Alexa, Google Home, Cortana, et.) [5, 2, 32, 54] which can be viewed as a combination of chatbots and goal-oriented dialogue systems. With the power of cloud computing and the internet of things technologies, these conversational AIs are changing our daily life and developing a 10 billion pounds worth industry [48]. Here, we would like to introduce the architectures of existing statistical spoken dialogue systems (hereafter SDSs) and their main challenges to perform coherent conversations in an open domain.

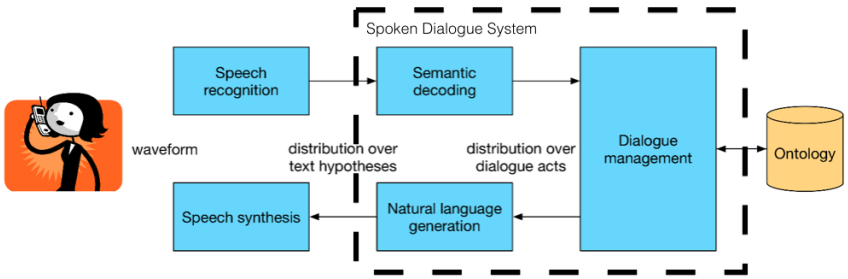


Figure 1.4: Architecture of a Spoken Dialogue System [29].

Current SDSs have three key components which are shown as Figure 1.4. A semantics decoder decodes the meaning in utterances into dialogue acts which describe the current intention of the user, for example, confirm(food=Korean). Then, a dialogue manager, which usually consists of a belief tracker and a policy network, will keep track of the belief states and produce a dialogue act as the response. The dialogue manager treats dialogue generation as a partially observable Markov decision process (POMDP) [82, 87, 86], which explicitly models the uncertainty natural of human conversations with Bayesian methods. Such a framework provides robustness against the errors created by speech recognition. Later a semantic encoder (natural language generator) will map this act back into a natural language response.

However, keeping track of the belief states under such a framework presents a tough challenge. Exact model representation is infeasible due to the limitation of computational complexity [87]. Carefully constructed approximation and additional independent assumption are needed. For example, if we are willing to find an expensive 5-star hotel in the city centre, in order to make the inference practical, we have to assume that there is no relation between a hotel is expensive and it is a 5-star hotel. It is not true. In addition, as shown in Figure 1.4, predefined ontology is needed in order to keep

track of the dialogue state. This makes the existing dialogue system infeasible to perform conversations in an open-domain. One research direction would be building an SDS which can support a natural conversation about any topic within a wide coverage Knowledge Graph (KG). This forms the definition of an open-domain spoken dialogue system [68].

1.2 Thesis Outline

In this thesis, we mainly focus on creating a framework in which neural models could learn linguistic patterns within extended coherent dialogues.

Firstly, we explore how to build such a framework by designing, collecting and annotating a conversational dataset. This dataset consists of extended coherent dialogues of questions and responses. These dialogues should describe the relations within a knowledge graph with clear links between the phrases in the questions to nodes in a knowledge graph. Secondly, we analyse the collected dataset statistically in order to investigate how human behave differently under different context in different domain. Later, we will introduce a proposed neural model to model these linguistic patterns and discuss how the existing dialogue systems could benefit from this model.

We start off by providing background knowledge on available corpora to train dialogue models and how to evaluate these dialogue models in Chapter 2. In Chapter 3 and Chapter 4, we will describe the design and creation of our framework in detail and perform analysis to figure out the common patterns within these dialogues. Chapter 5 will introduce our proposed neural models and methods in order to model coherent dialogues with neural networks. We will briefly summarize our findings and mention some further work in Chapter 6.

Chapter 2

Background

In this chapter, we investigate the existing resources and frameworks to build a neural dialogue model and summarize the popular evaluation metrics for dialogue modelling tasks.

2.1 Corpora for Training Dialogue Models

Humans perform dialogues with different purposes, under different situations and in different media. There is a vast amount of data available documenting human communications. Some of these data is collected as corpora, and these corpora can be classified into different categories. Each category has different characteristics, utilities and applications. In order to make the most of the available resources, we discuss the important distinctions between each type of dialogue corpora: whether a corpus features constrained or unconstrained dialogues; whether a corpus is written, spoken; whether a corpus features human-human interactions or human-machine interactions. Afterwards, we give a summary of current publicly available corpora.

2.1.1 Chit-chat or Goal-oriented Corpora

People use different languages for different purposes. We may want to call an airliner to book a ticket or chat with our friends to share our brilliant ideas. The purpose and motivation of our conversation will have a significant influence on the language we use. Similarly, how a dialogue corpus is collected can also have a significant influence on the model we built from. Some conversations are relatively causal, informal and unconstrained, which we usually call them chit-chats. There are many chit-chat corpora, and most of them are aiming to mimic spontaneous and unplanned spoken interactions between humans. Consequently, they are also referred to as Spontaneous (Unconstrained) Corpora [72]. Another large proportion of the existing corpora focus on a particular topic or intend to solve a specific task. In such situations, the task or topic is pre-specified, and participants are discouraged from deviating from the topic. We refer to these as Task-Oriented or Constrained Dialogue Corpora.

Dialogues in spontaneous corpora usually have richer grammar and vocabularies, which

reflect our daily communication well. However, such corpora are usually in plain text with less informative annotations. Consequently, people usually train models with end-to-end methods using those corpora. Such a setting could reduce the interpretability of models' reasoning and pose additional challenges to models' evaluation.

For constrained dialogues, some experimental conditions in which dialogues were collected can result in unnatural behaviours. Consequently, dialogues in goal-oriented corpora usually do not bear a close resemblance to our daily communication and do not correlate well to the actual typical dialogue patterns in natural conversations. Therefore, all experimental conditions have to be carefully designed to make sure that the conversations have the desired properties we are looking for. Not only the experiment instruction but also many other factors (e.g. the demographics of the participants [1, 87]) could significantly influence the data we collected. It is generally hard to do before we run the experiment. Designing a constrained corpus is particularly challenging.

2.1.2 Written, Spoken and Multi-modal Corpora

Another important distinction between dialogue corpora is whether participants interact through written language, spoken language, or in a multi-modal setting (e.g. using both speech and visual modalities). Written language and spoken language differ from each other and have substantially different linguistic properties. Spoken language tends to be informal, containing less information per utterance and have more pronouns than written language [14, 8].

Similarly, dialogues involving visual and other modalities differ from dialogues without these modalities [71, 25]. When a visual modality (e.g. body language, eye contact, hand motions) is available, such non-verbal information could influence the distributions of many linguistic properties significantly [30, 51, 18, 15, 21]. Our corpus is solely text-based. We will justify this design decision in Chapter 3. In addition, our corpus could be extended to a multi-modal setting, which would be a matter of future work.

2.1.3 Human-Human or Human-Machine Corpora

Another salient distinction between dialogue corpora resides in the types of interlocutors. Conversations between two humans are different from conversations involving machines.

This distinction is important because current computer dialogue systems exhibit very different traits than human-human dialogues [24]. Machines are significantly constrained and normally produce less changeable dialogues. These systems do not produce the same distribution of possible responses as humans do under the same conditions. In addition, human-machine dialogues contain different types of errors and the turn-taking in these dialogues is much more predictable [82]. Consequently, in order to learn natural conversations, it is not sensible to learn from human-machine corpora, as the trained dialogue system would simply learn to approximate the policy of the system in the corpus [72].

Another possible experiment setting is called Wizard-of-Oz experiments [9, 60, 12, 26]. In Wizard-of-Oz experiments, a human thinks he/she is speaking to a machine dialogue system, but a human is controlling the dialogue system. This is very similar to Turing's test but in reversed order. Conversations in such setting are likely to have similar distributions as the situation when the dialogue system has a nearly human-level performance. This can encourage machines to learn from human intelligence and achieve comparable performance when the behaviour of another human participant is not seriously different.

2.1.4 Available Dialogue Corpora

Covering all available corpora for training dialogue models is infeasible; therefore, we would only focus on three popular categories: the human-human spontaneous dataset, the human-human constrained dataset and the human-machine dataset. We would like to give examples which are well-known or directly related to this project. Table 2.1 and Table 2.2 summarize those available corpora and compare the key features between them.

Human-human Spontaneous Corpora

People talk to each other for entertainment. Here, we classify all corpora that participants talk without a predefined objective as spontaneous (chit-chat) corpora. Typically, the domain of these corpora would be unconstrained.

The majority of such spontaneous corpora are spoken. The Switchboard dataset [31] is one of the most influential spontaneous spoken corpora. This corpus consists of approximately 2,500 dialogue transcriptions from phone calls with about 500 different speakers. A computer program introduces a topic for discussion between two participants and records the conversation. There are approximately 70 casual topics in which 50 of them are frequently used. The corpus was initially designed for automatic speech recognition (hereafter ASR) task, and it has also been used for other tasks. Another remarkable corpus is the British National Corpus [43] (BNC), which consists of about 10 million words of spoken conversations and 100 million words of written articles. The spoken part was collected in a broad range of contexts ranging from formal business or government meetings to casual phone conversations, and the written samples were extracted from regional and national newspapers, published research journals and many other publications. BNC covers a variety of sources and is representative to a wide cross-section of British English from the late twentieth century. The corpus is annotated with part-of-speech (POS) tags for every word. The decent topics coverage makes BNC be a very good resource as a general-purpose spoken dialogue dataset.

Instead of using spoken language, there are also numerous corpora based on different media. With the rapid development of the Internet industry, a tremendous amount of conversations are recorded. For example, the online chatroom is a good source of mining conversations. NPS Internet Chatroom Conversations Corpus [27], which contains 10,567 English utterances gathered from age-specific chatrooms of different online chat platforms. This corpus is one of the first corpora use the Internet to be its media of communication. After that, several Internet-based corpora have also been collected

from a variety of sources. Instead of chatrooms, lots of conversations also happen in micro-blogs or discussion forums. Twitter Corpus [66] contains 1.3 million post-reply pairs extracted from Twitter, which was originally constructed for unsupervised dialogue acts modelling. Many other Twitter corpora with larger-scale have also been collected. The Twitter Triples Corpus [75] is one such example, which consists of 127 million context-message-response triples. Not only for English but also many other languages, large micro-blogging corpora have been created, such as, the Sina Weibo Corpus [73], which contains 4.5 million post-reply pairs in Chinese.

Corpora mined from the Internet are usually larger than corpora collected from other types of media by an order of magnitude or more. This is a significant advantage for deep learning methods. However, these corpora collected from the Internet has an enormous amount of typos, slang, and abbreviations. Particularly for Twitter, due to the 140-character limitation, tweets are often very short and compressed [72]. Another challenge is that Twitter conversations often rely on implicit contexts, for example, recent public events outside the conversation. In order to model such conversations, a dialogue model must be able to infer such implicit contexts by referencing some form of external knowledge source. This is a particularly tricky task.

Type	Name	Year	Source	Number of Dialogues	Number of Words	Annotation
Human-human spontaneous corpora	Switchboard [31]	1992	Spoken	2,400	3M	
	British National Corpus [43]	1992	Spoken/Written	854	10M/100M	Part of speech tags.
	NPS Chat Corpus [27]	2007	Online Chat	704	100M	Dialogue acts and part of speech tags.
	Twitter Corpus [66]	2010	Microblog	1.3M	125M	
	Twitter Triple Corpus [75]	2015	Microblog	4232	65K	
	DailyDialog [47]	2017	Online Chat	13K	15M	Emotion of speakers.
	The Cambridge and Nottingham Corpus of Discourse in English [53]	1998	Spoken	-	5M	Relationship between speakers and interaction type.
	D64 Multimodal Conversation Corpus [58]	2013	Video	2	70K	Physical head, torso and arm motion.
	Topical Chat [33]	2019	Online Chat	10784	4M	Links between plain text knowledge and conversation.

Table 2.1: Example Human-human Spontaneous Corpora.

Human-Human Goal-oriented dataset

For many corpora, the topic of the conversation specified beforehand, and participants are discouraged from deviating. Most of the participants in those conversations are aiming to solve a specific task or help the other participants to solve the task. This may introduce biases which influence the distributions of the linguistic properties and human behaviours in the conversations. As a result, any restrictions introduced to the conversations should be carefully considered. On the other hand, imposing restrictions can also bring immense benefits. A goal-driven framework makes it possible to apply reinforcement learning to explicitly optimise some desired properties (e.g. coherence, vocabulary diversity) which are generally impossible to learn with supervised learning [40]. The success rate of the task or the goal could also be an effective extrinsic evaluation metric which we will discuss in a later section. Here, we give some typical tasks for the existing goal-oriented corpora.

Several corpora focus on task planning or path-finding through the collaboration of two participants. In these corpora, one person acts as the decision maker and the other acts as the observer. The observer usually helps to decision-maker to achieve his/her goals. A well-known example corpus is the HCRC Map Task Corpus [4], which has unscripted, task-oriented dialogue transcriptions. In this corpus, each participant must collaboratively reproduce a route from one participant's map to the map of another participant via natural language conversions.

Another popular scenario is persuasion or debate. A common task for a specific participant in the conversation can be to convince another participant of some opinion or topic. Generally, these corpora are annotated with the outcome of the debate, for example, how convinced an audience is of the argument after the conversation. The Intelligence Squared Debate Dataset [88] covers the "Intelligence Squared" Oxford-style debates with a variety of topics. However, for each session of debate, the topic of that debate is predefined and constrained. The outcome of the debate is provided (how many of the audience members were for the given proposal or against, before and after the debate).

Another popular task is to perform question-answering (hereafter QA) grounding on external knowledge. Many researchers are aiming to create a mini-game where two players ask and answer sequential questions in turns. In each turn, an interlocutor has to ask questions about a predefined fact, or he/she could decide what to ask in a given domain. The other interlocutor has to answer the questions based on the conversation history accordingly. The majority of the corpora in this category provide an external knowledge base for each participant instead of grounding the conversations on participants' own knowledge. This could reduce the implicit bias introduced by different participants.

GuessWhat [22] is a two-player game where the goal is to identify an object in a complex visual scene by asking a sequence of yes/no questions. There are 150k human-human dialogues collected which contain 821k questions together with Boolean answers. A human participant has to ask sequential questions related to a given image in order to find the location of an object. The location is predefined and only visible to one of the participants (the one answers the questions). This mini-game forms a common

type of tasks, which is usually called a visual question answering (VQA) task. There are many efforts towards this direction [76, 74, 65, 89, 22, 19, 20], in which they try to model visually grounded conversations using different machine learning paradigms (including RL) in a question answering setting. However, the imperfect performance of the image encoding network leads to a poor RL policy, which is the bottleneck of the whole system. Instead of learning the best conversation strategy, the model learns to find the most effective protocol by utilising the strength of the answering network. This protocol will clearly not produce natural conversations.

Instead of grounding conversations on images, Reddy et al. [65] proposed a conversational question answering challenge (CoQA) which contains question answering conversations in natural language grounded on a short paragraph with its supporting evidence. However, the top models on the leader board surpass human performance. Even if we ignore the sequence labelling nature of the task, which humans may underperform, it is clear that even the state of the art neural model would not achieve human intelligence on these kinds of natural language understanding tasks. The results above demonstrate that whatever representations these models learn, they are fundamentally different from what we are expecting. This emphasises the importance of the interpretability of a network's reasoning. We need a framework in which we can perform a controlled evaluation because only then we can justify its decisions.

To the best of our knowledge, there is no publicly available corpus suitable for training coherent open-domain dialogue models. Such a corpus should make transparent when an utterance makes a coherent contribution to its context vs when it does not. The corpus should contain both coherent conversations as well as incoherent conversations. It should be collected (much like the HCRC map task corpus was) from competent human language users, in an unscripted fashion.

We are aiming to achieve this by manipulating what the humans must address next in their conversation via a knowledge graph (this controls whether the next utterance is coherent or not), but not controlling how they say and what they have to say (so it is the human's choice on whether to use anaphoric expressions or elided constructions or not). This bears many similarities to the way the HCRC map task corpus was collected, but there are also differences. Instead of using maps, we decide to use structured knowledge graphs (to present the task but also to control the coherence of the next dialogue move). Both of their tasks and ours incorporate asymmetry, but different kinds, among the dialogue participants. Participants in the HCRC map task could see maps with or without annotated routines. However, our task will provide the same knowledge graphs with the same markings, but they are aiming to achieve different goals using those graphs.

In Chapter 3, we will discuss our task design in detail and justify our design decisions.

Type	Name	Year	Source	Number of Dialogues	Number of Words	Description
Human-human Goal-oriented Corpora	HCRC Map Task Corpus [4]	1991	Spoken	128	147K	Dialogues from HLAP Task in which speakers must collaborate verbally to reproduce on one participants map a route printed on the others.
	Intelligence Squared Debate Dataset [88]	2016	Debates	854	10M	Various topics in Oxford-style debates, each constrained to one subject. Audience opinions provided pre- and post-debates.
	GuessWhat [22]	2017	Mini Game	160K	4M	Identify an object in a complex visual scene by asking a sequence of yes/no questions.
	Complex Sequential Question Answering [69]	2010	Semi-auto Generated	169K		Sequential question answering utterance generated based on human defined template.
	CoQA [65]	2019	QA Chat	8K		Question answering conversation grounded on short paragraph. Labeled with text spans as supporting evidence.
	QuAC [17]	2018	QA Chat	14K	5.6M	Question answering conversation grounded on wikipeida text. Labeled with text spans as supporting evidence. Annotated with dialogue acts.

Table 2.2: Example Human-human Goal-oriented Corpora

Human-Machine Corpora

Most of the human-machine corpora are goal-oriented. Humans usually talk to machines in order to accomplish specific tasks or enquiry information. The ATIS (Air Travel Information System) Pilot Corpus [34] is one of the first human-machine corpora. It consists of conversations between human participants and a travel-type booking system, secretly operated by humans. This dataset contains 1041 utterances. The Carnegie Mellon Communicator Corpus [7] also contains human-machine interactions with a travel booking system. It is a medium-sized dataset of interactions with a real-time system providing flight information, hotel information, and car rentals. The user's comments are recorded after each dialogue.

Instead of making queries or making reservations, people also talk to machines for other purposes. The DIALOG mathematical proof dataset [84] is a Wizard-of-Oz dataset involving an automated tutoring system that attempts to advise students on proving mathematical theorems. In the dataset, a system gives heuristics that provide clues when students come up with incorrect answers. There are only 66 dialogues in the dataset which consists of a conglomeration of text-based interactions with the system, as well as think-aloud audio and video footage recorded by the users as they interacted with the system.

2.2 Dialogue Model Evaluation

Evaluating dialogue models is one of the most challenging aspects of building dialogue systems. Dialogue systems are generally evaluated by humans. However, human evaluation is very time consuming and expensive. Although user satisfaction is our ultimate goal for building a dialogue system, it is often necessary to optimise the performance on some automatic metrics for multiple times prior to release. It is inefficient and almost infeasible to run experiments with real users during the whole development circle. In this section, we investigate some commonly used approaches for dialogue system evaluation and discuss the pros and cons of each approach. Those approaches are roughly classified into three categories: Intrinsic Evaluation, Extrinsic Evaluation and Human Evaluation.

2.2.1 Intrinsic Evaluation

Word Overlap Metrics

As we mentioned in Chapter 1, modelling dialogues could be views as a machine translation task. Consequently, we may borrow machine translation evaluation metrics, such as BLEU scores [59], to evaluate dialogue models. BLEU score could be applied to calculate the word overlap between a machine-generated utterance and the actual next utterance in the conversation. However, for assessing dialogue system responses, such a measure has been reported not to correlate well with human judgment [50]. One significant issue is that natural language is ambiguous. For a given utterance, there will be a vast amount of valid responses. These responses may not share the same word at all. In this case, systems producing diverse and "interesting" responses, will score

poorly with BLEU score. Sordoni et al. [75] have demonstrated that human responses may be scored poorly according to word overlap metrics.

Word Perplexity

For probabilistic language models, word perplexity is a well-established performance metric [6, 55]. Word perplexity has also been suggested for evaluating generative dialogue models [61]. Perplexity explicitly measures the probability that the model will generate the gold next utterance, given the prior context of the conversation. Low perplexity demonstrates good performance. The probabilistic nature of perplexity can potentially evade the exact matching issue of BLEU and consider multiple possible valid responses.

2.2.2 Extrinsic Evaluation

For goal-oriented dialogue systems, we could directly use the performance of the downstream tasks to reflect the performance of our dialogue models. Many previous works [76, 74, 65, 89, 22, 19, 20] have adapted this method. They typically focus on goal-related performance criteria, such as success rate, accuracy, precision and recall.

Another paradigm is adversarial evaluation [11, 41, 46], inspired by the Turing test. The idea is to train a "Turing-like" evaluation classifier to distinguish between human-generated responses and machine-generated responses. One can narrow the number of possible responses to a predefined list, and ask the model to select the most appropriate response from this list. The list includes the actual next response of the conversation (the desired prediction), and the other entries (false positives) are sampled from elsewhere in the corpus. The more successful a response generation system is at fooling this evaluator, the better the system. There are several advantages to this task: it is easy to interpret, and its difficulty can be adjusted by changing the number of false responses. However, there are drawbacks. Since the other candidate answers are sampled from elsewhere in the corpus, there is a chance that these also represent reasonable responses given the context.

2.2.3 Human Evaluation

There are many aspects in dialogue systems which have no well established automatic evaluation metrics. For example, it would be complicated to come up with an algorithm to determine how natural or appropriate a machine response is. Consequently, human evaluation is required. Crowd-sourcing platforms, such as Amazon Mechanical Turk [3], are wildly used for dialogue system evaluation. Many toolkits for setting up such an experiment have been developed [42]. However, we should pay attention that evaluations using paid subjects may also lead to biased results [87]. Consequently, evaluation experiments should be carefully designed before production.

Chapter 3

Extended Dialogues Grounded on Knowledge Graph: A Corpus

In this chapter, we design a novel corpus contains extended coherent dialogues of questions and responses. These dialogues describe the relations within a knowledge graph with clear links between the phrases in the questions to nodes in the knowledge graph. An example knowledge graph looks like Figure 3.1. The purpose of this corpus is to perform statistical analysis about various linguistic patterns within coherent dialogues, thereby creating a framework for neural models to learn how to parse these patterns and produce more coherent responses. The main focus for this project is to investigate when the appearance of an elided construction is natural and conveys the intended content. We firstly introduce our framework design. Then, we discuss how this framework is implemented and how the corpus is collected. Finally, we compare our work with other existing corpora and show the advantage of our framework in achieving the above goal.

3.1 Corpus Design

In order to create a framework for neural networks to model various linguistic patterns within a coherent dialogue and ultimately build a dialogue system which can perform natural conversations in open-domain, there are some important requirements of our framework:

1. The dialogues should be coherent and natural.
2. The framework should be able to provide strong supervision signals to train our neural models.
3. The framework could be potentially adapted to open-domain.

These are the design principles for our corpus. To achieve the above goal and collect natural conversations, human interlocutors are necessary. In our corpus, human-human conversations are collected instead of human-machine or machine-generated conversations. Williams and Young [82] have argued that, under equivalent circumstances,

machines are producing a different distribution of possible responses than humans. As we mentioned in Chapter 1, small divergences or misunderstanding of the context will lead to a completely incoherent and unnatural conversation. Consequently, it is very hard to produce coherent conversations when machines are involved as interlocutors. The conversations are collected via a text-based online chatroom in order to reduce the biases introduced by automatic speech recognition [82] as well as remain the informal natural of chit-chat conversations. All the conversations are goal-oriented, in fact, topic-oriented, and the participants are discouraged from deviating from the topic. As we mentioned in Chapter 2, this facilitates the training (apply reinforcement learning) and testing (extrinsic evaluation) for our model development. Participants are instructed to perform question-answering tasks about a knowledge graph. Consequently, we could use common question-answering evaluation metrics such as precision, recall and F1 score to be our external evaluation metrics. The sequences of questions and answers are predefined. By doing this, we could manually create different contexts where we hypothesis a linguistic pattern (e.g. ellipsis) could possible appear or not, and test our hypothesis statistically. If our hypothesis is significant, we may build probabilistic models with neural networks to model this pattern.

As outlined above, our corpus should be considered as a human-human goal-oriented written corpus. In the rest of this section, we give a concise definition of our task and a brief description of our corpus. Then, we introduce three important aspects of our corpus design and justify our decisions.

3.1.1 Task Definition

Our task pairs up two participants: a Questioner and an Answerer. They discuss the relations between entities within a sampled knowledge graph. A knowledge graph is a directed graph, which the nodes in the graph represent entities in real-world; the edges represent relations among these entities; the arrow on an edge indicates the direction of the relation. Figure 3.1 is an example knowledge graph talking about Alan Turing.

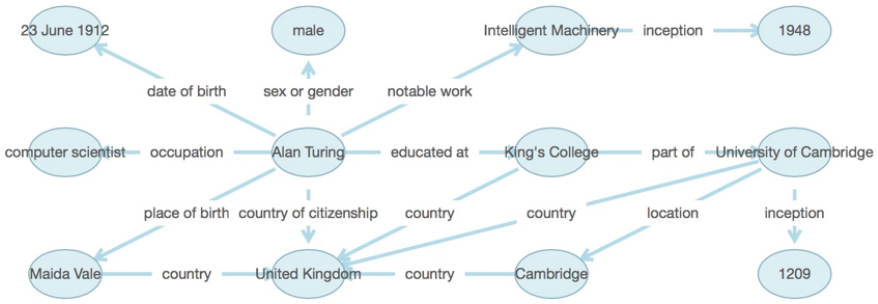


Figure 3.1: An Example Knowledge Graph

The Questioner and the Answerer will talk in turns. They have the freedom to choose how to phrase their sentences and keep the conversation natural, although, what they are talking about is constrained. In each turn, there will be a unique highlighted (red) relation in the graph together with a highlighted (green) node in this edge. The participants should only talk about the unique highlighted edge and node in each turn.

The Questioner is the initiator of the conversation. In order to start a conversation, the Questioner should ask a question based on the relation (edge) in red. The node marked in green should be the answer to the question. The Questioner will send the natural language question together with the same graph with the same marking to the Answerer. After the Answerer received the question, he/she has to answer the question according to the marked graph, and the conversation history so far. Then the Answerer will send the response to the Questioner, and the marking on the graph will be updated. Now the two participants have to start the next turn. One important fact is the conversation history so far is always visible to both participants. Consequently, their utterances will always depend on the context of the on-going conversation. A more detailed instruction is described in Appendix B.

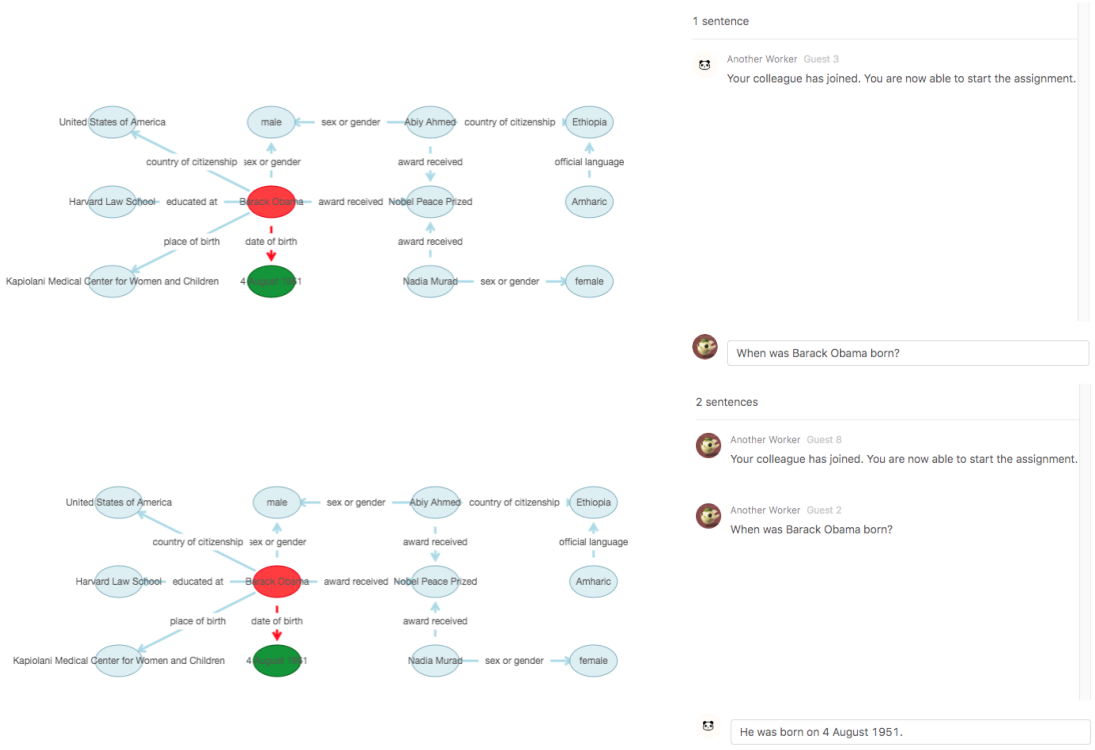


Figure 3.2: Question Answering Interface for Questioner(above) and Answerer(below).

A conversation will be automatically terminated after six turns. In a conversation, the markings (highlights) will be different, but the knowledge graph remains the same. During the experiment, participants are unable to see each other. There is no non-verbal commutation between the two participants, and they are paid for this task.

3.1.2 Corpus Description

Our corpus consists of 240 coherent extended conversations collected via crowd-sourcing. Each conversation contains six turns, and each turn is composed of a natural language question and a natural language answer to that question. All utterances in a conversation are correlated to its prior context and the grounded knowledge graph and linked to the marking on the knowledge graph of that turn. Figure 3.3 is an example data point in our corpus.

(2) a. Q: What date was Alan Turing born?
b. A: He was born on 23 June, 1912
c. Q: Where?
d. A: Hmm... I believe in Maida Vale
e. Q: And do you happen to know what his occupation was?
f. A: He was what's known as a computer scientist
g. Q: Do you know any of his notable work?
h. A: Yeah, intelligent machinery
i. Q: where was he educated?
j. A: In King's College- this dude was a smarty!
k. Q: What university is that part of?
l. A: University of Cambridge- the best!!

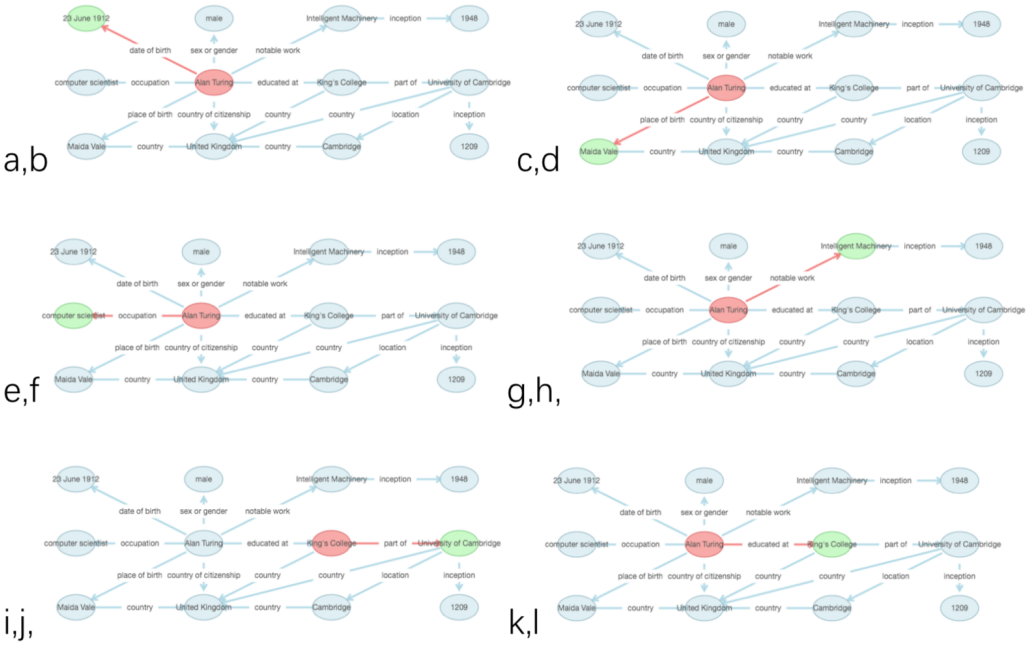


Figure 3.3: An Example Collected Conversation and its Grounded Knowledge Graph.

All the knowledge graphs are sampled from WikiData [78], and each graph has 12 nodes. The sequences of highlighted relations and nodes were designed to control for the coherence of the current question in its discourse context so that the corpus as a whole incorporates variation in the extent to which the topic of the current question shifts from the previous topics of prior questions and answers. Human annotators predefine this sequence of marking. Some demographic information of the participants, such as age, gender and nationalities, is collected under their permission. In addition, all the data entries are annotated with labels in Table 3.1 and Table 3.2 which will be discussed later in this section.

3.1.3 Important Aspects of the Corpus

Knowledge Graph as External Knowledge

Instead of letting the participants ask questions solely based on their prior common knowledge, external information as grounded knowledge is provided. Different people have different interpretations of the world. Conversations based on participants' common knowledge often rely on implicit contexts about the participants. This brings additional challenges to produce dialogue state representations.

As we mentioned in Chapter 2, using images or text paragraphs for modeling open domain coherent dialogues in a question answering setting [76, 74, 65, 89, 22, 19, 20, 65] has limitations. In our framework, we use knowledge graphs as our external knowledge base. There are many advantages to using knowledge graphs as external knowledge:


```
{
  "edges": [
    {
      "id": "P27Q7251Q145",
      "label": "country of citizenship",
      "source": "Q7251",
      "target": "Q145",
      "wikiID": "P27"
    }...
  ],
  "nodes": [
    {
      "id": "Q145",
      "label": "United Kingdom"
    },
    {
      "id": "Q7251",
      "label": "Alan Turing"
    }...
  ],
  "path": [
    {
      "answer": "Q145",
      "edge": "P27Q7251Q145"
    }
  ]
}
```


Figure 3.4: An Example Knowledge Graph in JSON Format

1. **Knowledge graphs have two modalities.** As illustrated in Figure 3.4 and Figure 3.1, a knowledge graph can be stored as structured data in the computer memory or visualized as an image graph. These two representations convey exactly the same information with different modalities. This provides both machine-readability and human-readability. By doing this, human participants do not underperform during the task and machines could learn effectively through the strong supervision signals it provided.

2. **Knowledge graphs are unambiguous.** Different people have different ways to interpret a sentence and label a text span. This makes the evaluation of the text-based question-answering task very challenging. As we mentioned above,

Term	Definition
Gaps	If there is no overlap between the entities in relations between the previous turn and the current turn, then we label the current turn with 'Gap' label.
Links	When two turns are talking about the same entity, we call the following, turn is linked to the previous turn if there is a strong correlation between, their grounded relationships.
Continuous	A conversation with no gap is called continuous.
Discontinuous	A conversation with one or more gaps is called discontinuous.

Table 3.1: Definition for Graph Annotations.

humans usually underperform in such tasks. Unlike plain text, knowledge graphs are unambiguous. Accurate evaluation is possible under this framework.

3. **Knowledge graphs are structured.** Knowledge graphs are structured in the sense that the relations between entities are easily interpretable. This makes it extremely easy to control the topics and topic shifts of the conversations.

4. **Knowledge graphs are open domain.** There are many existing knowledge graph datasets [78, 10]. These knowledge graphs cover a wide range of topics which could be considered as open-domain.

5. **Knowledge graphs are well-studied.** There are many prior works related to knowledge graphs, such as embedding [49]. Also, a knowledge graph is a directed graph. There are many algorithms (e.g. depth-first search) we could use for our study.

As outlined above, knowledge graphs are good resources of external knowledge in our task. There are two popularly knowledge graphs which are widely used in literature. We decide to use WikiData [78] instead of Freebase [10]. WikiData has an operating API. Researchers could directly query the API instead of dealing with raw data which can be hundreds of Gigabit.

Natural Conversation Grounded on Knowledge Graph

For each conversation, there will be six relations and answer nodes highlighted in sequence. The sequence of the highlighted relations together with the highlighted nodes, which we call a path, is predefined in order to control the topic shifts of each conversation. Different patterns of topic shifts (paths) will influence the appearances of various linguistic patterns. In Table 3.1, we would like to introduce some terminologies before we investigate those patterns. Figure 3.5 and Figure 3.6 are example sequences of gaps and links.

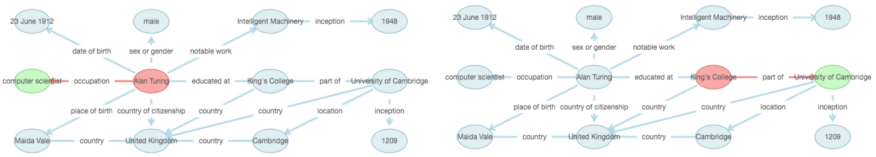


Figure 3.5: An Example Marking Sequence of Gaps.

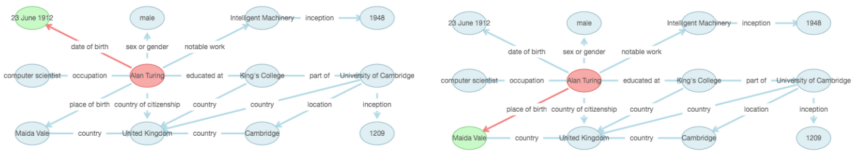


Figure 3.6: An Example Marking Sequence of Links.

We sampled ten sub-graphs from WikiData in ten distinct topics. For each sub-graph, we generate four paths with a different number of gaps and links in each path. By doing this, we would like to investigate when humans tend to produce various linguistic patterns, particularly elided constructions, in an extended coherent dialogue.

Annotations for Linguistic Patterns

Each utterance in the corpus is tagged with various linguistic patterns by human annotators. There are eight tags in our corpus. Four of them are considered to be class labels. Consequently, tagging those tags is a classification task which the annotators see the whole utterance and annotate the tags to the whole sentence. The rest four tags are sequence labels, which these tags are labelled to a text span of the original utterance. Table 3.2 summarizes the details about our tag set. All annotators are tagging the corpus according to the annotation schema in Appendix A.

Here, we give dialogue extracts for each tag described in Table 3.2. These utterances are sampled from our corpus. For each ellipsis, we give the full underlying meaning of that utterance in the following parentheses.

1. **Sluicing**

   (3) a. Q: What date was Alan Turing born?
   b. A: He was born on 23 June, 1912
   c. Q: Where? (=Where was Alan Turning born?)
   d. A: Hmm... I believe in Maida Vale

   Question (3c) should be tagged with "Sluicing" in our corpus. In this utterance, the Questioner is interrogating about the place of birth of Alan Turing. However, without the context, the full underlying meaning of this wh-expression is uninterpretable.

2. **Anaphora**

   (4) a. Q: Who founded SoftBank?
   b. A: That was Masayoshi Son. (The person who founded SoftBank is Masayoshi Son.)

   Answer (4b) is considered to be Anaphora because we could not fully interpret the meaning of this utterance solely based on itself without the prior context. The appearance of a pronoun usually indicates an occurrence of Anaphora constructions.

Type	Tag	Definition
Classification	Sluicing	Sluicing is a type of ellipsis that occurs in both direct and indirect interrogative clauses. The ellipsis is introduced by a wh-expression, whereby in most cases, everything except the wh-expression is elided from the clause.
	Anaphora	Anaphora is the use of an expression whose interpretation depends upon another expression in context (its antecedent or postcedent). In our schema, anaphora is the use of an expression that depends specifically upon an antecedent expression.
	Short Answer	Short Answer (= answer fragments) is a type of ellipsis that occurs in answers to questions. In our schema, we define all answer utterances which is not a fully grammatical sentence to be short answer.
	Error	Error tag should only be annotated when there is a disagreement between the grounded knowledge graph and the utterances.
Labelling	Additional Information Outside Graph	If parts of an utterance introduce additional information into the conversation which is not mentioned in the grounded knowledge graph, we call this part additional information outside the knowledge graph.
	Additional Information Within Graph	If parts of an utterance introduce additional information into the conversation which is mentioned in the graph and such information is not highlighted, we call them additional information within the knowledge graph.
	Conversation Control	Conversation Control label should be tagged to any pieces of utterance which functionally act as a connector to make the whole conversation more natural and coherent.
	Totally Irrelevant	Totally Irrelevant tag should be used when pieces of utterance is totally irrelevant to the question answering conversation. By deleting such pieces of information, we will not influence the coherence of the conversation or the meaning of the utterance. In addition, this tag should only be considered if the part of utterance is not labelled as other labels. In other words, labels, such as Conversation Control, will have higher precedence.

Table 3.2: Definition of our Tag Set.

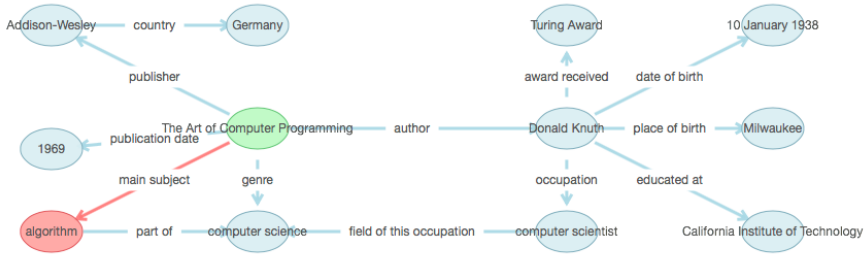
3. **Short Answer**

(5) a. Q: What operating system does the iphone have?
b. A: iOS. (Iphone has the iOS operating system.)

Answer (5b) is an example of Short Answer which the ellipsis occurs in answer to Question (5a). The object noun is an answer fragment, and the elided material could be inferred from the context.

4. **Error**

Assume the highlighted knowledge graph is given below.

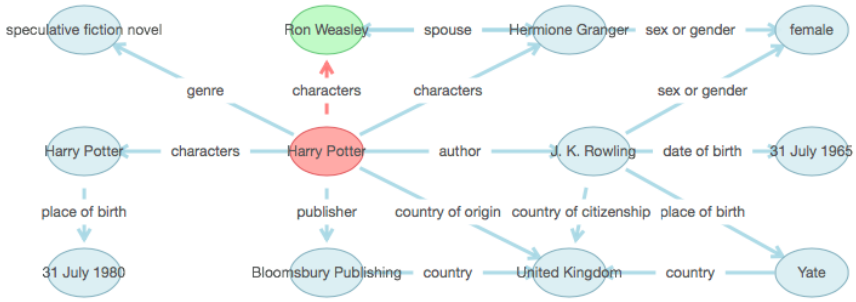


(6) a. Q: What is the main subject of “The Art of Programming”?

Question (7a) is tagged with Error tag. There is a disagreement between the highlighted knowledge graph and the utterance. The highlighted relation instructs the participant to ask a question which the answer of that question should be “The Art of Computer Programming”.

5. **Additional Information Outside Graph**

Assume the highlighted knowledge graph is given below.

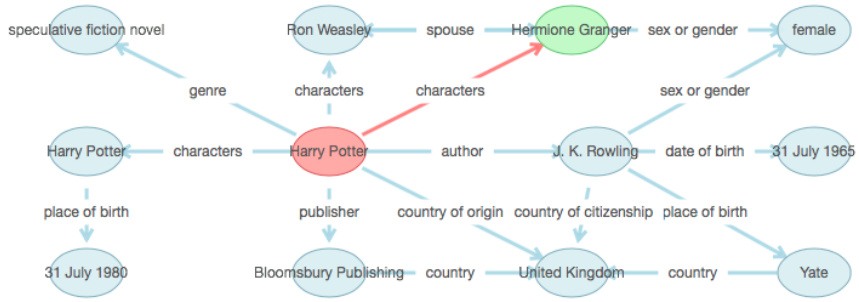


(7) a. Q: Who is the main red head character of the series?

The underlining utterance is considered as additional information outside the knowledge graph because there is no evidence that Ron Weasley has red hair.

6. **Additional Information Within Graph**

Assume the highlighted knowledge graph is given below.



(8) a. Q: Nice. So she's another main character, you said?

b. A: Yeah she is a character of the film, one of the 3 main ones i would say

The underlining utterance is labelled as additional information within the knowledge graph. The highlighted relation does not indicate that there are three main characters in Harry Potter. However, we could infer this knowledge from the given knowledge graph.

7. **Conversation Control**

(9) a. Q: Let's do a 180 here. Is there anything following the CPU cache?

The underlining utterance is labelled with Conversation Control label. This piece of the sentence does not provide additional information. However, it maintains the coherence of the overall conversation. It is a signal to the Answerer that there will be a topic shift in this question.

8. **Totally Irrelevant**

(10) a. A: San Francisco - where my daddy's from!.

The underlining utterance is labelled with Totally Irrelevant label. This information provided is not related to our predefined task. Deleting the underlining utterance will not influence the logic and the coherence of the overall dialogue.

3.2 Implementation

This section introduces the toolkits have been developed and used during this project. In addition, we briefly discuss the technical difficulties during the project and their solutions.

3.2.1 Sampling from Knowledge Graph

In order to sample sub-graphs from WikiData, a Java toolkit named KnowledgeGraph-Client is developed.

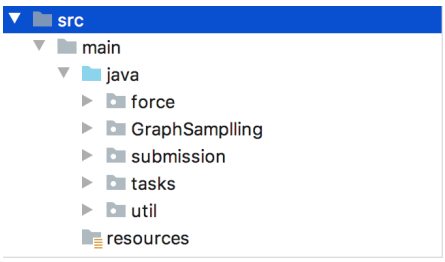


Figure 3.7: Structure of KnowledgeGraphClient.

In order to interact with WikiData, a Java package name WikiData Toolkit [81] is used. This package provides APIs to query WikiData for entities (nodes in Figure 3.1) and relations related to it.

A sub-graph is sampled via a depth-first search algorithm. We predefined a set of "interesting relations" which we only explore a connecting node if the relation connecting that node and the current node is in the set. We do not want to include relations which are rarely mentioned in our daily life. For example, we would rarely talk about relations like Obama is an instance of human and human is a subclass of animal. We start the search at a predefined node, for example, Obama(Q76). Q76 is the unique WikiData ID for Obama. We potentially could traversal all nodes which are connected to Obama; however, we set the maximum depth of the search to be four. In other words, we only explore nodes which are 4 connections away from the starting node. After the search, we would have a set of edges and nodes which will form a directed graph. In order to visualise the graph, we tried to apply the force-directed algorithm to reduce the overlap among these nodes.

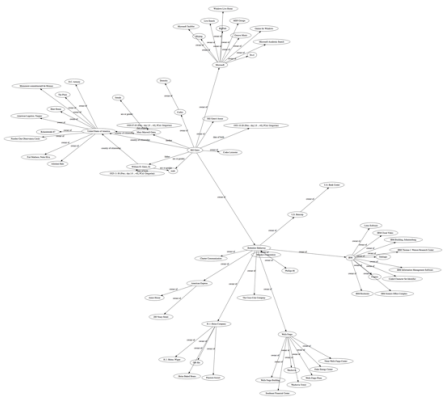


Figure 3.8: An Example Force-Directed Graph.

However, such a graph is not very human interpretable and visible. We decide to follow a semi-automatic approach which we manually select nodes and edges from the above graph and directly give the coordinate for each node. In addition, the paths are also manually generated. By this approach, we generate 40 graphs in 10 distinct topics.

Predicate	CEO
Subject	Sundar Pichai
Object	Google
Parent Predicate	designation
Domain	person
Range	organization

Table 3.3: An Example Set of Keywords Constructed from the Triple (Sundar Pichai, CEO, Google) [39].

Model	BLEU Score
BiRNN-Large-Reported	50.14
BiRNN-Small-Experiment	38.45
BiRNN-Large-Experiment	49.72

Table 3.4: Experiment Result for Generating Questions [39].

3.2.2 Generating Questions from Knowledge Graph

Instead of collecting dialogues from human participants, Indurthi et al. [39] have tried to generate questions from a given knowledge graph automatically. They use a subset of keywords to generate a natural language question that has a unique answer. They treat this subset of keywords as a sequence and use an encoder-decoder model to generate natural questions from it.

We re-implemented their experiment in a similar setting, and the experiment results are shown in Table 3.4. The larger model is a bi-directional RNN containing one hidden layer of 1000 units. We also build a small model, which uses a bi-directional RNN containing one hidden layer of 500 units due to the limitation of our computation power. However, their method faces domain-adaption problem, which could not produce grammatical sentences with out-domain data. Consequently, this dataset requires human intelligence and crowd-sourcing is required.

3.2.3 Web-interface for Data Collection

In order to collect this corpus via crowd-sourcing, we built a web-interface together with a web-server from scratch. A large proportion of our effort for this project has been spent developing this software. There are two major challenges during the development: 1) This is a production system, which real users have to interact with. Consequently, the system has to be thoroughly tested. 2) Sometimes, human behaviours are unpredictable. It is very challenging to design and implement a robust system that could handle unexpected human behaviours.

As shown in Figure 3.9, if a new user participates in our experiment, he/she will be shown a detailed instruction about the experiment. If the participant decides to take part in the experiment, he/she has to sign a consent form according to the data protection regulation. If the user agrees with the terms and conditions of the experiment, after registering an account, the user could see the web-page for our task. More detailed

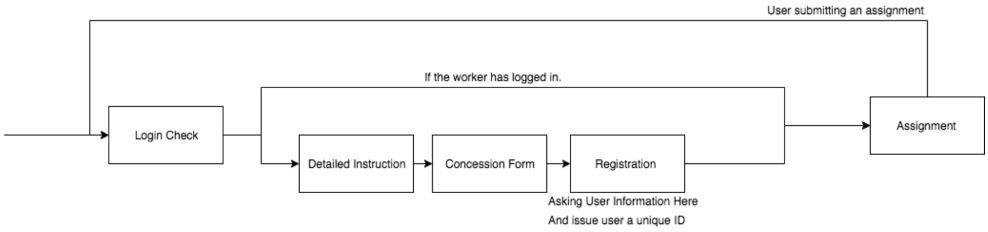


Figure 3.9: The Workflow of the Web-interface.

Package	Purpose
NodeJS [28]	Writing web-server.
MongoDB [38]	Database for storing user information and collected data.
ReactJS [37]	Build user interface.
AntD [35]	Components for user interface.
G6 [36]	Rendering knowledge graphs in front end.
Socket.IO [64]	Writing the socket server for online chatting.

Table 3.5: Packages Imported for Developing the Web-interface.

workflow and screenshots of this web-interface are shown in Appendix B. In order to develop this web-interface, we have used many packages in Table 3.5. Besides, we keep privacy and data security in mind during our design, implementation and experiment. The server is running over https with due efforts to comply with the data protection regulations.

3.2.4 Tools for Annotating the Dataset

In order to annotate the dataset, a toolkit named doccano [57] is used. This is an easy to use web-application which provides features for the sequence to sequence tasks, the classification tasks and the sequence labelling tasks. Many other annotation tools have also been explored. However, most of them are professional and complicated to set up.

In addition, we develop a tool for post-processing and parsing our corpus, which is frequently used during our data analysis.

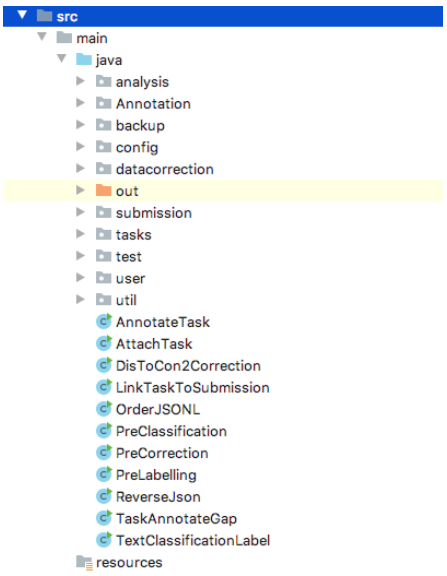


Figure 3.10: Toolkit for Post-processing and Parsing the Dataset.

3.3 Data Collection Procedure

In this section, we introduce how the corpus is collected. The whole data collection experiment runs for two months which is the most challenging and time-consuming part of this project. Totally 30 experiment sessions have been scheduled. The project is funded by Informatics Student Service, and each participant will be paid a 5 pounds Amazon Voucher for his/her work. These are several experiment constraints which make organizing and running this experiment extremely hard:

1. Each experiment session needs two participants at the same time.
2. Only native speakers could participate in this experiment.
3. Building a robust system that multiple users could interact with simultaneously is hard.

In this section, we discuss our data collection procedure in the following order: Advertising the Experiment and Scheduling the Experiment. Finally, we would like to give some advice for future researchers who are aiming to collect a dialogue corpus on a larger scale.

Advertising the Experiment

In order to recruit enough participants for this project, advertising this experiment is crucial. We have tried several methods:

1. Emailing the whole Informatics students.
2. Pinning leaflets around the campus (see Appendix C for details).
3. Posting advertisements on social media (Facebook).

Emailing students is the most effective approach by which we recruit most of the participants. An alternative approach is to deploy this experiment to crowd-sourcing platforms, such as Amazon Mechanic Turk [3]. Hundreds of participants could be recruited overnight if the experiment is reasonably paid.

Scheduling the Experiment

The most challenging part of the data collection and even the whole project is to schedule these experiment sessions. It is very difficult to pair up the participants and find a suitable time for both of them. We have tried to make a google form and ask participants to select their available time slots. We also tried to send a group email to all the interested participants and ask them if they are available for a fixed time session. Because people do not check their emails regularly, it takes time for our participants to reply. Consequently, all of the above approaches do not work well for our experiment.

Finally, we start following this procedural, and it works reasonably well.

1. Participants have to be registered on a list if they are interested in this project.
2. Build an online survey and ask participants roughly when are they available next week.
3. Pair up participants who are available on the same date.
4. Ask permission to share their emails with other participants.
5. Send this pair of participants a group email and ask them to find a suitable time themselves.
6. Confirm the attendance one day before the experiment session and send them a detailed instruction for that session.

However, when the scale of the corpus goes larger (thousands of conversations), we strongly encourage researchers to hire full-time in-house annotators, if the resource is available.

3.4 Data Annotation Procedure

The whole corpus is annotated by in-house annotators (my friends and I) according to an annotation schema (See Appendix A). The majority of the annotation work is done by myself. A subset of the corpus (10%) has been annotated twice by different annotators, in order to calculate the inter-annotator agreement. Cohen's Kappa is used for calculating the agreement between annotators [13].

3.5 Comparison with Previous Work

Our corpus contains extended natural conversations grounded on an external wide-coverage structured knowledge graph. At present, there is no such dataset existing (see Table 3.6), and this is for what our corpus mainly developed.

Dataset	Year	Multi-turn	Natural Conversation	Answer Type	External Knowledge Type
CoQA [65]	2018	✓	✓	Text Span	Text
CSQA [69]	2018	✓	✗	Text Answer	Knowledge Graph
SQuAD [63]	2016	✗	✗	Text Span	Wikipedia articles
SQuAD 2.0 [62]	2016	✗	✗	Text Span	Wikipedia articles
QuAC [17]	2018	✓	✗	Text Span	Wikipedia articles
FiloIT [74]	2017	✗	✗	Text Span	Images
GuessWhat [22]	2017	✓	✗	Object Position	Images
Our Dataset	2020	✓	✓	Natural Text Response	Knowledge Graph

Table 3.6: A Comparison between this Corpus with Existing Relevant Question Answering Corpora.

Chapter 4

Analysis and Discussion

In this chapter, we perform quantitative and qualitative analyses based on our corpus. We articulate several hypotheses and use observations in the annotated corpus to test these. We use the t-test to measure statistical significance.

4.1 Summary of the Corpus

In the previous chapter, we introduce a novel conversational question answering corpus. Our corpus contains 240 natural conversations, each has six question-answering turns, on ten distinct topics. These conversations are divided into two categories: continuous and discontinuous conversations. There are 140 continuous and 100 discontinuous conversations. Discontinuous conversations are additionally split into two subgroups, discontinuous1 (conversations with only one gap) and discontinuous2 (conversations with many gaps). There are 43 "dis1" conversations and 57 "dis2" conversations. Table 4.1 shows a detailed count for conversations of each topic in each category.

Table 4.2 shows basic statistics about our corpus. Figure 4.1 demonstrates the com-

Topics	WikiData ID	Continuous	Discontinuous1	Discontinuous2
Alan Turing	Q7251	13	6	5
JK Rowling	Q34660	14	6	6
Donald Knuth	Q17457	19	0	6
Obama	Q76	9	5	6
Berkshire Hathaway	Q217583	12	7	5
Softbank	Q201653	18	0	6
Apple	Q312	11	7	6
CPU	Q5300	19	0	7
London	Q84	12	5	6
Tianjin	Q1173	13	6	5

Table 4.1: Detailed Topics Count for the Corpus.

Number of Dialogues	240
Number of Questions	1.4K
Number of Answers	1.4K
Number of Words	1.8K
Number of Tokens	15.2K

Table 4.2: Counts of the Corpus.

mon relations which are asked in the conversations. This also reflects what types of questions are asked in the corpus.

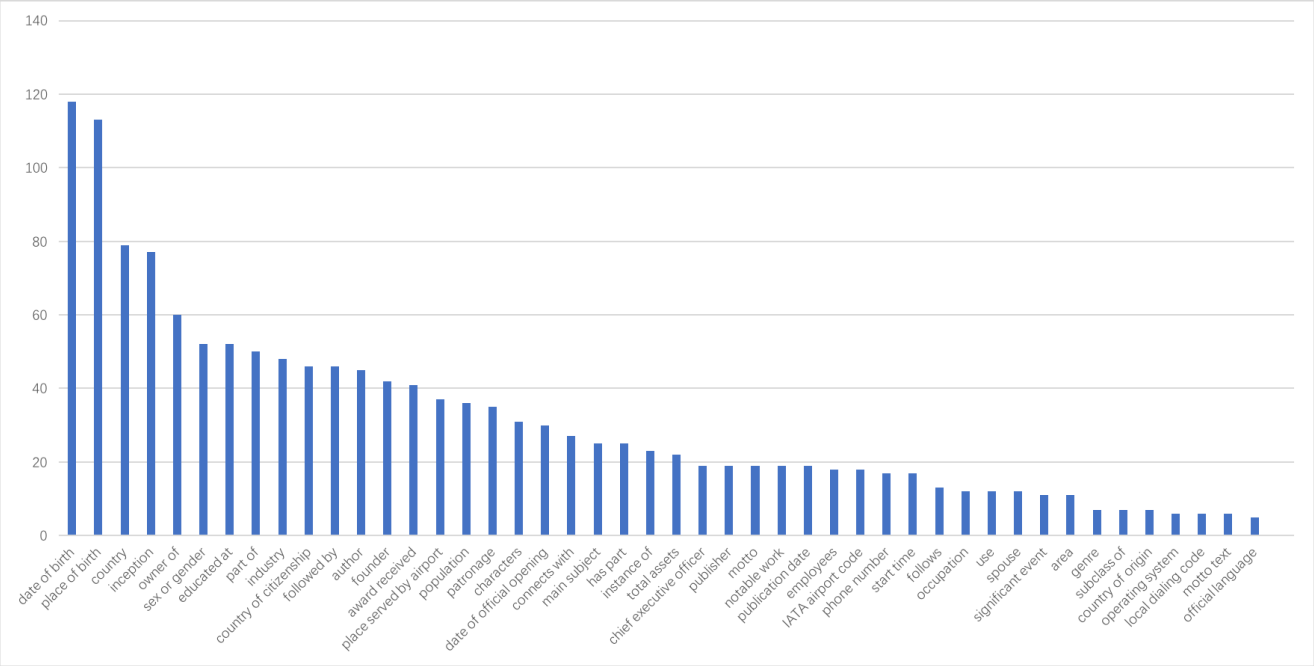


Figure 4.1: Counts of Common Relations in the Corpus.

Inter-Annotator Agreement of the Corpus

The inter-annotator agreement is a measure of how well two (or more) annotators can make the same annotation decision for a certain category. Cohen's Kappa is used here for calculating the agreement score between two annotators [13]. The higher the agreement score for a tag, the more confident the tag is valid. High agreement scores are desirable properties for a corpus which demonstrate the good quality of its annotations.

We calculate the agreement score for all tags in Table 3.2. According to Table 4.3, Anaphora and Short Answer have an almost perfect agreement. Labels such as "Additional Information Outside Graph" and "Conversation Control" have a lower Kappa score. This behaviour is expected, and there are several reasons:

1. According to the annotation schema (Appendix A), the criteria for tags, such as Anaphora, is less ambiguous. A different interpretation of the semantics of a

Tag	Cohen's Kappa	Agreement
Anaphora	0.9476	Almost perfect agreement
Short Answer	0.9370	Almost perfect agreement
Error	0.0	Slight agreement
Sluicing	0.0	Slight agreement
Totally Irrelevant	0.0	Slight agreement
Additional information outside graph	0.3405	Fair agreement
Additional information within graph	-0.0084(invalid)	Poor agreement
Conversation Control	0.6060	Moderate agreement

Table 4.3: Inter-annotator Agreement of the Corpus

sentence usually does not influence the judgment of that tag.

2. There are obvious heuristics for some tags, such as Anaphora and sluicing. For example, the appearance of a pronoun may indicate the appearance of an anaphora and the appearance of a wh-word indicates the possibility of a sluicing.
3. Tags, like Anphora, can usually be determined by the grammars instead of the semantics of a sentence.
4. Annotating Anphora is a classification task, which is easier than a sequence labelling task.
5. Annotators have to make inference about the conversation context in order to make a decision for tags like "Additional Information Within Graph".

In addition, we notice that the Kappa score is 0.0 for many tags. Some tags are extremely sparse in the corpus, and many rarely appear within the 10% of the dataset.

Demographics of the Participants

There are a total of 17 native English speakers involved in our experiment. Seven of them are female, and ten of them are male. The majority of the participants (11) are from the United Kingdom, four participants are from Singapore, and two participants are from the United States. All participants are students from the University of Edinburgh, and all of them are between 18 and 24 years old.

4.2 Statistical Analysis

In this section, we perform a statistical analysis on the corpus. We additionally make some hypotheses according to our observations. The significance of each hypothesis is tested via t-test.

Condition		
Total Number of Utterances	2880	
Tags	Count	# per Utterance
Anaphora	1197	0.4156
Short Answer	902	0.3132
Error	11	0.0038
Sluicing	27	0.0094
Additional information outside graph	50	0.0174
Additional information within graph	26	0.0090
Conversation Control	181	0.0628
Totally Irrelevant	26	0.0090

Table 4.4: Statistics of Tags for the Corpus

Table 4.4 shows the count of each tag in the whole corpus. In this table, the "Condition" is empty. It means all the utterances in this corpus are considered in this table. If the "Condition" is "Question", we only consider all the utterances which are questions in this table. In this table, it shows how many utterances satisfy the given "Condition". It also shows that how many utterances contain each tag and the proportion of the utterances contain that tag in all the utterances satisfy the "Condition".

According to Table 4.4, that there are 2880 utterances in the dataset, and 1197 of them contain Anaphora tag which consists of 41.56% of these utterances.

Condition	Continuous	
Total Number of Utterances	1680	
Tags	Count	# per Utterance
Anaphora	737	0.4387
Short Answer	529	0.3149
Error	6	0.0036
Sluicing	25	0.0149
Additional information outside graph	41	0.0185
Additional information within graph	9	0.0054
Conversation Control	104	0.0619
Totally Irrelevant	13	0.0077

Table 4.5: Statistics of Tags for Continuous Conversations.

Condition	Discontinuous	
Total Number of Utterances	1200	
Tags	Count	# per Utterance
Anaphora	460	0.3833
Short Answer	373	0.3108
Error	5	0.0042
Sluicing	2	0.0017
Additional information outside graph	19	0.0158
Additional information within graph	17	0.0142
Conversation Control	77	0.0642
Totally Irrelevant	13	0.0108

Table 4.6: Statistics of Tags for Discontinuous Conversations.

Condition	Discontinuous1	
Total Number of Utterances	516	
Tags	Count	# per Utterance
Anaphora	242	0.4690
Short Answer	153	0.2965
Error	0	0.0000
Sluicing	1	0.0019
Additional information outside graph	14	0.0271
Additional information within graph	9	0.0174
Conversation Control	51	0.0988
Totally Irrelevant	5	0.0097

Table 4.7: Statistics of Tags for Discontinuous 1 Conversations.

Condition	Discontinuous2	
Total Number of Utterances	684	
Tags	Count	# per Utterance
Anaphora	218	0.3187
Short Answer	220	0.3216
Error	5	0.0073
Sluicing	1	0.0015
Additional information outside graph	5	0.0073
Additional information within graph	8	0.0117
Conversation Control	26	0.0380
Totally Irrelevant	8	0.0017

Table 4.8: Statistics of Tags for Discontinuous 2 Conversations.

According to our experiment design in Chapter 3, gaps represent non-connected topic shifts. The more non-connected topic shifts a conversation has, the less coherent it is. In our corpus, discontinuous1 (dis1) conversation has only one gap per conversation.

Dis1 conversations represent conversations which the underlying logic is less coherent but still logistical. In contrast, discontinuous 2 (dis2) conversations have many gaps, which the underlying logic is totally random.

According to Table 4.5 - 4.8, we have the following observations and hypotheses.

Observation 4.1. *Continuity and coherence influence linguistic patterns.*

Hypothesis 4.1. *In our corpus, answers are always related to their questions. Gaps will have little effect on answers.*

In order to prove the above hypothesis, we additionally check the difference of the counts of tags between questions and answers for both continuous and discontinuous conversations. Table 4.9 - 4.12 demonstrate the results. These results show obvious differences between the appearance of tags in questions and do not make sharp differences in answers. This confirms the above hypothesis. Consequently, for the following statistics, we focus on questions instead of answers.

Condition	Continuous Question	
Total Number of Utterances	840	
Tags	Count	# per Utterance
Anaphora	465	0.5536
Short Answer	0	0.0000
Error	3	0.0036
Sluicing	25	0.0298
Additional information outside graph	14	0.0167
Additional information within graph	2	0.0024
Conversation Control	70	0.0833
Totally Irrelevant	6	0.0071

Table 4.9: Statistics of Tags for Continuous Questions.

Condition	Discontinuous Question	
Total Number of Utterances	600	
Tags	Count	# per Utterance
Anaphora	251	0.4183
Short Answer	0	0.0000
Error	4	0.0067
Sluicing	2	0.0033
Additional information outside graph	9	0.0150
Additional information within graph	7	0.0117
Conversation Control	53	0.0833
Totally Irrelevant	4	0.0067

Table 4.10: Statistics of Tags for Discontinuous Questions.

Condition	Continuous Answer	
Total Number of Utterances	840	
Tags	Count	# per Utterance
Anaphora	272	0.3238
Short Answer	529	0.6298
Error	3	0.0036
Sluicing	0	0.0000
Additional information outside graph	17	0.0202
Additional information within graph	7	0.0083
Conversation Control	34	0.0405
Totally Irrelevant	7	0.0150

Table 4.11: Statistics of Tags for Continuous Answers.

Condition	Discontinuous Answer	
Total Number of Utterances	600	
Tags	Count	# per Utterance
Anaphora	209	0.3483
Short Answer	373	0.6217
Error	1	0.0017
Sluicing	0	0.0000
Additional information outside graph	10	0.0167
Additional information within graph	10	0.0167
Conversation Control	24	0.0400
Totally Irrelevant	9	0.0067

Table 4.12: Statistics of Tags for Discontinuous Answers.

According to Table 4.9 and Table 4.10, we find that:

Observation 4.2. *Coherence and continuity of a conversation have a significant influence on Anaphora and Sluicing.*

Observation 4.3. *To answers, Short Answer and Anaphora are not very sensitive to the coherence of the conversation. It may also because the answer is always related to the question.*

Based on the above observations, we could see a decreasing number of Anaphora and Sluicing when the conversations are less coherent. We are additionally interested if there is a determining factor in such conversations which controls these linguistic patterns. We believe that the gaps in discontinuous conversations majorly influence the

appearance of Anaphora and Sluicing. Intuitively, gaps represent non-connected topic shifts, which usually introduce additional entities into the conversation context. The appearance of these new entities makes the co-reference resolution more difficult because determining which entity the pronoun refers to in the context is more ambiguous. A human may decide to use less elliptical constructions in this case. We compare the appearance of Anaphora and Sluicing between utterances without gaps and utterances with gaps in Table 4.13 and Table 4.14.

Condition	Question without Gap	
Total Number of Utterances	993	
Tags	Count	# per Utterance
Anaphora	675	0.6798
Short Answer	0	0.0000
Error	3	0.0030
Sluicing	26	0.0262
Additional information outside graph	16	0.0161
Additional information within graph	4	0.0040
Conversation Control	99	0.0997
Totally Irrelevant	8	0.0081

Table 4.13: Statistics of Tags for Questions without Gaps.

Condition	Question with Gap	
Total Number of Utterances	447	
Tags	Count	# per Utterance
Anaphora	41	0.0917
Short Answer	0	0.0000
Error	4	0.0089
Sluicing	1	0.0022
Additional information outside graph	7	0.0157
Additional information within graph	5	0.0112
Conversation Control	24	0.0537
Totally Irrelevant	2	0.0081

Table 4.14: Statistics of Tags for Questions with Gaps.

According to Table 4.13 and Table 4.14, we have the following observation.

Observation 4.4. *Gaps have a profound influence on Anaphora and Sluicing. The number of Anaphora and Sluicing increases in questions without gaps.*

Based on this observation, we make two hypotheses.

Hypothesis 4.2. *Gaps influence Anaphora. An utterance is more likely to contain Anaphora if there is no gap between this utterance and its previous utterance.*

The t-test score for this hypothesis is 5.3805×10^{-106}. This score rejects the null hypothesis, which there is a very little probability (5.3805×10^{-106}), the above hypothesis happens by chance. We prove that the above statement is statistically significant. We also have a similar statement for Sluicing.

Hypothesis 4.3. *Gaps influence Sluicing. An utterance is more likely to contain Sluicing if there is no gap between this utterance and its previous utterance.*

The t-test score for this hypothesis is 4.1702×10^{-8}, which also demonstrates the significance of the above hypothesis.

Condition	Linked Question without Gap	
Total Number of Utterances	98	
Tags	Count	# per Utterance
Anaphora	63	0.6429
Short Answer	0	0.0000
Error	0	0.0000
Sluicing	21	0.2143
Additional information outside graph	2	0.0204
Additional information within graph	0	0.0000
Conversation Control	13	0.1327
Totally Irrelevant	0	0.0000

Table 4.15: Statistics of Tags for Linked Questions without Gaps.

Condition	Unlinked Question without Gap	
Total Number of Utterances	895	
Tags	Count	# per Utterance
Anaphora	612	0.6838
Short Answer	0	0.0000
Error	3	0.0034
Sluicing	5	0.0056
Additional information outside graph	14	0.0156
Additional information within graph	4	0.0045
Conversation Control	86	0.0961
Totally Irrelevant	8	0.0089

Table 4.16: Statistics of Tags for Unlinked Questions without Gaps.

Similarly, we are interested in the effect of links to these tags. From Table 4.15 and Table 4.16, we observe that:

Observation 4.5. *Links have an obvious influence on Sluicing. The utterance with a link to its previous utterance is likely to contain a Sluicing construction. Links do not have an obvious influence on Anaphora.*

Based on this observation, we make two hypothesises:

Hypothesis 4.4. *Links influence the appearance of Sluicing. The utterance with a link to its previous utterance is likely to contain a Sluicing construction.*

The t-test score for this hypothesis is 1.3983×10^{-9}. The none hypothesis is rejected, which demonstrates the significance of the above hypothesis.

We would also prove that there is no obvious relation between links and Anaphora.

Hypothesis 4.5. *Links affect the appearance of Anaphora.*

This hypothesis achieves a t-test score of 0.4226. It means that there is 42.26% probability the pattern we observed in the hypothesis happens by chance. In other words, there is no statistically sufficient evidence to show there is a correlation between the appearance of links and the appearance of Anaphora.

4.3 Qualitative Analysis

There are many interesting patterns in the corpus. However, we are unable to show them statistically because of the limited size of the corpus. In this section, we give numerous examples of participants' utterances. Based on these examples, we discuss our observations and the implication about the challenges to build a coherent open-domain dialogue system.

(11) a. Q: Which country was Alan Turing a citizen of?
b. A: He was a citizen of the United Kingdom.
c. Q: By birth or immigration?
d. A: By birth; he was born in Maida Vale.

The underlining question (11c) asks about the place of birth of Alan Turing. Without the context, a human would typically ask this question by using the interrogative "where" or explicitly motioning the "place of birth". However, because the previous question (context) has mentioned his citizenship, our participant decides to phrase the question as (11c). This utterance will be extremely hard for an artificial dialogue system to parse, understand and produce a valid coherent response to. There is no overlap between the words in the utterance and the tokens in its underlying relations in the knowledge graph ("place of birth" in Figure 3.1). Actually, the problem is even harder. The question (11c) does not explicitly ask the place of birth. Consequently, a human has to make a decent inference to work out the underlying meaning of this question.

(12) a. Q: And what is Mr Turing's place of birth?
b. A: He was born in Maida Vale.

c. Q: Notable work of Turing is 'Intelligent Machinery', when was this incepted?

d. A: 1984.

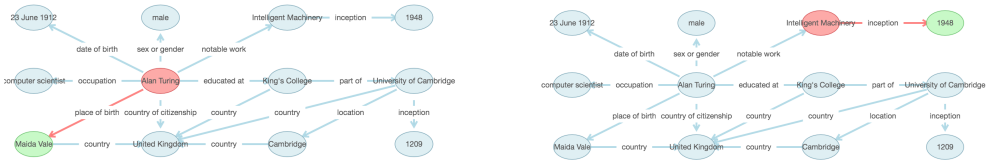


Figure 4.2: The Underlying Marking for Example (12). Left Figure for Turn a,b and Right Figure for Turn c,d.

As demonstrated in Figure 4.2, there is a gap in turn c,d in Example (12). Humans try to introduce additional information to connect gaps between each turn. These utterances could be information within the conversation context or information within the grounded graph but not explicitly mentioned. However, people sometimes give up when the conversations are totally incoherent.

(13) a. Q: Hey bro, I've been living under a rock my whole life... do you know who wrote Harry Potter?

b. A: J.K Rowling did man.

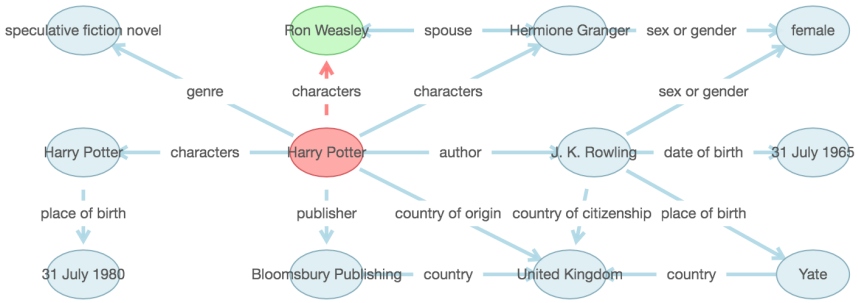


Figure 4.3: Example Knowledge Graph of JK Rowling.

(14) a. Q: So, where did she write Harry Potter?

b. A: She wrote harry potter in the United Kingdom of Great Britain and northern Ireland, mostly in the muggle part.

c. Q: Did any of her main characters NOT grow up in the muggle world?

d. A: A few grew up with Ron Weasley.

Example (13) is an example that human introduces additional utterances to improve the coherence. In Example (14), the human participant uses additional information outside the grounded knowledge graph 4.3 to disambiguate the potential answers in the graph. It is worth mentioning, in our instructions, human participants are not encouraged to add additional utterances to the dialogues, but they decided to do so in order to improve the coherence of the overall conversations.

All the examples and patterns we mentioned above are extremely challenging and unachievable for the current state of the art systems. Building coherent dialogue systems involves many fundamental problems of artificial intelligence and will remain a central topic for the research communities.

Chapter 5

Modelling Coherent Dialogues Using Neural Networks

In order to model the coherent dialogues in this corpus, we proposed a novel neural architecture. However, this model is not implemented due to the limitation of the scale of this project. This corpus only contains 240 conversations. As we argued in Chapter 2, this is insufficient to train such a complicated neural network. In this chapter, we introduce our proposed framework to train a neural dialogue model if the resources are available in the future.

5.1 Hierarchical Attentional Encoder with Context Windows

In order to encode coherent dialogues, we need to model long-distance dependencies not only between words in one utterance but also between sentences within the discourse. For example, in order to model an elided construction, our model has to infer the omission from the longer prior context of the conversation, because such information may not appear in the current utterance. To model that, we propose a hierarchical attentional encoder which is inspired by Xing et al. [85]. Our model design is shown in Figure 5.1. We extend their model with context windows which the size of the windows is a predefined hyper-parameter to control how many previous turns we are modelling.

We would like to use BERT [23] to provide a contextual embedding for each word. Additional turn embedding is used to denote the turn and type of a sentence. The concatenated embedding is fed into a self-attentional layer which models the dependencies within the context window. Additional sentence-level attention will be used to model the dependencies between context windows. In addition, we will adapt a multimodal deep neural network architecture to produce the dialogue state representation for downstream tasks.

To predict the correct node to answer a given question, we could represent a prediction as a quadruple, such as, (Subject Node, Relation, Object Node, Position of the

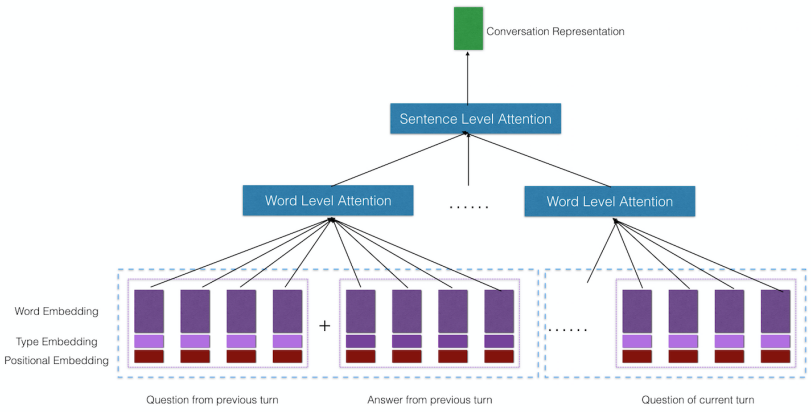


Figure 5.1: Hierarchical Attentional Encoder with Context Window.

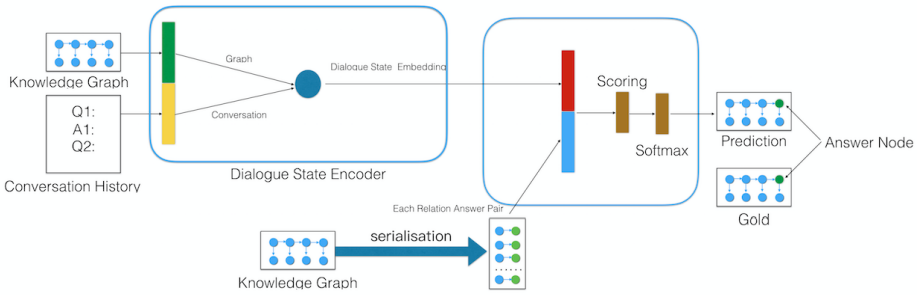


Figure 5.2: The Multi-modal Question Answering Architecture.

Answer). An example of the prediction could be (Barack Obama, born, 1967, object) which corresponds to the question "When was Obama born?". We permute all the possible combinations within the graph. Then we use a feed-forward neural network together with a softmax layer to produce a multinomial distribution over all the possible answers. Then the prediction is compared with the gold answer, and we could calculate the reward or loss to train our model. The detailed architecture is shown in Figure 5.2.

5.2 Generate Coherent Dialogues as a Markov Decision Process

As we argued in Chapter 2, training a dialogue model with supervised learning is not explicitly optimised to produce a coherent response. We propose to cast the response generation task as an RL task. More specifically, we consider the questioner and the answerer as part of the RL agent environment. According to Williams and Young [82], a goal-driven dialogue system could be modelled as a partially observable Markov decision process. Inspired by Strub et al. [76], we would like to describe our response generation task as following:

We define the state S_t as the state of the dialogue at step t. An action u_t corresponds to selecting a new word in the vocabulary V. The transition to the next state depends on

the selected action:

- If we generate the end of sentence token . , the ongoing response is terminated and the next question is sampled from the questioner.
- Otherwise the new word is appended to the ongoing answer.

A dialogue is started with an initial question and finished after a fixed number of turns. A reward is defined for every state-action pair.

5.3 Extrinsic Evaluation of Dialogue Models

We will use precision and recall, which are well-established metrics for question answering tasks, as our extrinsic evaluation metrics. In our framework, high precision and recall demonstrate a good performance of modelling the linguistic patterns within a coherent dialogue.

Chapter 6

Conclusion

6.1 Summary

In this project, we first briefly introduce the history of and progress in dialogue systems in terms of rule-based systems, corpus-based systems and the architecture of state of the art spoken dialogue systems. Then we discuss the important distinctions between different dialogue corpora: whether a corpus features constrained or unconstrained dialogues; whether a corpus is written, spoken; whether a corpus features human-human interactions or human-machine interactions. Afterwards, we review the publicly available corpora for training dialogue systems. We also investigate Intrinsic Evaluation, Extrinsic Evaluation and Human Evaluation methods for dialogue system evaluation and discuss the pros and cons of each approach.

The motivation for this project is to build a coherent open-domain dialogue system with neural networks. To the best of our knowledge, there is no publicly available corpus suitable for training coherent open-domain dialogue models. To achieve this goal, we design and collect a novel corpus which contains extended coherent dialogues of questions and responses. These dialogues describe the relations within a knowledge graph with clear links between the phrases in the questions to nodes in a knowledge graph. The purpose of this corpus is to address the gap in our understanding of the interaction between topic shifts on the one hand and the natural use of Anaphora and elided constructions. This corpus has been collected via crowd-sourcing. The dialogue collection experiment was running for two months which 30 experiment sessions have been organized. In order to prepare for this experiment, several toolkits have been developed.

The main focus for this project is to investigate when the appearance of an elided construction is natural and convey the intended content. A detailed statistical analysis has been conducted on this corpus. We find:

1. Gaps have an influence on Anaphora and Sluicing. An utterance is more likely to contain these patterns without a gap.

2. Links have a profound influence on Sluicing; however, links do not have a noticeable influence on Anaphora.

We prove the statistical significance of those hypotheses on our corpus. We also show how challenging and exciting to model coherent dialogues via examples in our corpus. Finally, building coherent dialogue systems will remain our research interests in the future.

6.2 Future Work

Due to the scale of this project and the limitation of the available resources, only 240 conversations have been collected in the corpus. In order to train decent dialogue models using deep neural networks, a larger corpus should be collected. For collecting a larger corpus, both the topics of the knowledge graphs and tags in our tag set could be further diversified. In addition, we could extend our corpus to a multi-model setting, which we could try to investigate how visual modality interacts with languages within coherent dialogues.

During this project, we have implemented a web-interface for this task-specific dialogue collection experiment. This interface is for a single purpose. A general-purpose toolkit for crowd-sourcing goal-orientated dialogue corpora should be developed. This toolkit should facilitate researchers to design, create and deploy such tasks into crowd-sourcing platforms, such as Amazon Mechanic Turk [3]. There are existing toolkits [42, 56]. However, it is not easy enough to create a similar experiment by only using these tools.

In Chapter 5, we proposed a hierarchical attentional neural architecture with context windows. Training such a neural model under our framework may improve the performance of current spoken dialogue systems. Although the current statistical SDSs do explicitly capture the dialogue history via belief tracking [29], they may not utilise rich linguistic information inside a coherent dialogue. Future studies could investigate how to integrate a conversation model into the semantic decoder and encoder of SDSs. This could enable the current SDSs to understand and produce more coherent dialogues. In addition, current SDSs require a predefined ontology in order to keep track of the belief state. Researchers could investigate whether we could connect an SDS to a wide coverage knowledge graph and build a dialogue system which can support natural conversations about any topics within the knowledge graph. Furthermore, Cheng et al. [16] have developed a transition-based neural semantic parser with a generic tree-generation algorithm. By integrating a conversation model with an existing neural semantic parser, we may enable it to perform semantic parsing over an extended coherent dialogue.

Bibliography

[1] Hua Ai, Antoine Raux, Dan Bohus, Maxine Eskenazi, and Diane Litman. Comparing spoken dialog corpora collected with recruited subjects versus real users. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 124–131, 2007.

[2] Amazon. Alexa internet, 2020.

[3] Amazon. Amazon mechanical turk, 2020.

[4] Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. The hcrc map task corpus. *Language and speech*, 34(4):351–366, 1991.

[5] Apple. Siri, 2020.

[6] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[7] Christina Bennett and Alexander I Rudnicky. The carnegie mellon communicator corpus. In *Seventh International Conference on Spoken Language Processing*, 2002.

[8] Douglas Biber, Edward Finegan, Terttu Nevalainen, and Leena Kahlas-Tarkka. Diachronic relations among speech-based and written registers in english. *Variation in English: Multi-dimensional studies*, pages 66–83, 2001.

[9] Dan Bohus and Alexander I Rudnicky. Sorry, i didn't catch that! In *Recent trends in discourse and dialogue*, pages 123–154. Springer, 2008.

[10] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, 2008.

[11] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

[12] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*, 2018.

[13] Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254, 1996.

[14] Ronald Carter and Michael McCarthy. *Cambridge grammar of English: a comprehensive guide; spoken and written English grammar and usage*. Ernst Klett Sprachen, 2006.

[15] Tanya L Chartrand and John A Bargh. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893, 1999.

[16] Jianpeng Cheng, Siva Reddy, Vijay Saraswat, and Mirella Lapata. Learning an executable neural semantic parser. *Computational Linguistics*, 45(1):59–94, 2019.

[17] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*, 2018.

[18] Roger M Cooper. The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 1974.

[19] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.

[20] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2951–2960, 2017.

[21] Iwan de Kok, Dirk Heylen, and Louis-Philippe Morency. Speaker-adaptive multimodal prediction model for listener responses. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 51–58, 2013.

[22] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[24] Christine Doran, John Aberdeen, Laurie Damianos, and Lynette Hirschman. Comparing several aspects of human-computer and human-human dialogues. In

Current and new directions in discourse and dialogue, pages 133–159. Springer, 2003.

[25] Starkey Duncan. Charles goodwin, conversational organization: Interaction between speakers and hearers. new york: Academic, 1981. pp. xii+ 195. *Language in Society*, 12(1):89–92, 1983.

[26] Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyag Gao, and Dilek Hakkani-Tur. Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*, 2019.

[27] Eric N Forsythand and Craig H Martell. Lexical and discourse analysis of online chat dialog. In *International Conference on Semantic Computing (ICSC 2007)*, pages 19–26. IEEE, 2007.

[28] OpenJS Foundation. Nodejs, 2020.

[29] Milica Gasic. Semantic decoding in dialogue systems. 2016.

[30] James J Gibson and Anne D Pick. Perception of another person's looking behavior. *The American journal of psychology*, 76(3):386–394, 1963.

[31] John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520. IEEE, 1992.

[32] Google. Meet google home., 2020.

[33] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895, 2019.

[34] Charles T Hemphill, John J Godfrey, and George R Doddington. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

[35] Alibaba Inc. Ant design: A ui design language and react ui library, 2020.

[36] Alibaba Inc. G6: A graph visualization framework in javascript, 2020.

[37] Facebook Inc. React: A javascript library for building user interfaces, 2020.

[38] MongoDB Inc. Mongodb: The database formodern applications, 2020.

[39] Sathish Reddy Indurthi, Dinesh Raghu, Mitesh M Khapra, and Sachindra Joshi. Generating natural language question-answer pairs from a knowledge graph using a rnn based question generation model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 376–385, 2017.

[40] Dan Jurafsky and James Martin. *Speech & language processing (Draft 3)*. 2019.

[41] Anjuli Kannan and Oriol Vinyals. Adversarial evaluation of dialogue models. *arXiv preprint arXiv:1701.08198*, 2017.

[42] Kyusong Lee, Tiancheng Zhao, Alan W Black, and Maxine Eskenazi. Dialcrowd: A toolkit for easy dialog system assessment. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 245–248, 2018.

[43] Geoffrey Neil Leech. 100 million words of english: the british national corpus (bnc). , 1992.

[44] Jiwei Li. *TEACHING MACHINES TO CONVERSEk*. PhD thesis, STANFORD UNIVERSITY, 2017.

[45] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

[46] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*, 2017.

[47] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017.

[48] Prescient Strategic Intelligence Private Limited. Conversational ai market is projected to surpass $15 billion by 2024: Ps intelligence. 2019.

[49] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[50] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.

[51] Catherine Lord and Marshall M Haith. The perception of eye contact. *Perception & Psychophysics*, 16(3):413–416, 1974.

[52] Ryan Thomas Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1):31–65, 2017.

[53] Michael McCarthy. *Spoken language and applied linguistics*. Ernst Klett Sprachen, 1998.

[54] Microsoft. Cortana - your personal productivity assistant, 2020.

[55] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

[56] Alexander H Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*, 2017.

[57] Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. doccano: Text annotation tool for human, 2018. Software available from https://github.com/doccano/doccano.

[58] Catharine Oertel, Fred Cummins, Jens Edlund, Petra Wagner, and Nick Campbell. D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, 7(1-2):19–28, 2013.

[59] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[60] Stefan Petrik. Wizard of oz experiments on speech dialogue systems. *Design and Realisation with a New Integrated Simulation Environment. Masters. Graz University of Technology, Graz. Institute of Signal Processing and Speech Communication*, 2004.

[61] Olivier Pietquin and Helen Hastie. A survey on metrics for the evaluation of user simulations. *The knowledge engineering review*, 28(1):59–73, 2013.

[62] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

[63] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[64] Guillermo Rauch. Socket.io 2.0 is here, 2020.

[65] Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

[66] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180. Association for Computational Linguistics, 2010.

[67] Alan Ritter, Colin Cherry, and William B Dolan. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pages 583–593. Association for Computational Linguistics, 2011.

[68] Lina Rojas. Open domain statistical spoken dialogue systems.

[69] Amrita Saha, Vardaan Pahuja, Mitesh M Khapra, Karthik Sankaranarayanan, and Sarath Chandar. Complex sequential question answering: Towards learning to

converse over linked question answer pairs with a knowledge graph. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[70] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[71] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*, 2015.

[72] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse*, 9(1):1–49, 2018.

[73] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*, 2015.

[74] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*, 2017.

[75] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*, 2015.

[76] Florian Strub, Harm De Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. End-to-end optimization of goal-driven and visually grounded dialogue systems. *arXiv preprint arXiv:1703.05423*, 2017.

[77] INTELLIGENCE BY AM TURING. Computing machinery and intelligence-am turing. *Mind*, 59(236):433, 1950.

[78] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledge base. 2014.

[79] Richard Wallace. Artificial linguistic internet computer entity (alice). *City*, 1995.

[80] Joseph Weizenbaum. Eliza, a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[81] Wikidata. Wikidata toolkit, 2020.

[82] Jason D Williams and Steve Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422, 2007.

[83] Terry Winograd. Shrdlu: a system for dialog. 1972.

[84] Magdalena Wolska, Quoc Bao Vo, Dimitra Tsovaltzi, Ivana Kruijff-Korbayová, Elena Karagjosova, Helmut Horacek, Armin Fiedler, and Christoph Benzmüller.

An annotated corpus of tutorial dialogs on mathematical theorem proving. In *LREC*, 2004.

[85] Chen Xing, Yu Wu, Wei Wu, Yalou Huang, and Ming Zhou. Hierarchical recurrent attention network for response generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[86] Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174, 2010.

[87] Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013.

[88] Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. Conversational flow in oxford-style debates. *arXiv preprint arXiv:1604.03114*, 2016.

[89] Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. A dataset for document grounded conversations. *arXiv preprint arXiv:1809.07358*, 2018.

Appendix A

Annotation Schema for this Corpus

Annotation Schema

Classification:

1. Sluicing

Definition: sluicing is a type of ellipsis that occurs in both direct and indirect interrogative clauses. The ellipsis is introduced by a wh-expression, whereby in most cases, everything except the wh-expression is elided from the clause.

Example:

Yes! It is an example of sluicing ☺:

> **Phoebe ate** something, but she doesn't know what. (=what she ate)
> Where? (=where was Alan Turning born)

No! It is not an example of sluicing ☹:

> Turing studied where? (It is an example of inverted sentence. It has a fully grammatical structure without any ellipsis. Although it has a similar format as sluicing, we should not label it as sluicing.)

2. Anaphora

Definition: anaphora is the use of an expression whose interpretation depends upon another expression in context (its antecedent or postcedent). In our schema, anaphora is the use of an expression that depends specifically upon an antecedent expression.

If you can not fully interpret the meaning of an utterance solely based on itself without the prior context, we call it anaphor.

Example:

Yes! It is an example of anaphora ☺:

> It opened in November 1939. (What is it?)
> Which place does this airport serve? (Which airport is this?)
> And what is the date of publication? (The publication of what?)
> Who is another character of Harry Potter? (Which character has been mentioned?)

No! It is not an example of anaphora ☹:

> What's the population of tianjin? (We can answer the question without prior context.)
> 24 February 1955. (Although the date is describing the date of birth of Steve Jobs according to prior context, we could still interpret this data without of reviewing previous context. In other words, there is no ambiguity arising in this sentence.

3. Short Answer:

Definition: short answer (= answer fragments) is a type of ellipsis that occurs in answers to questions. In our schema, we define all answer utterances which is not a fully grammatical sentence to be short answer.

Example:

Yes! It is an example of short answer ☺:

> Q: Who walked the dog? A: Tom walked the dog. **- Subject noun as answer fragment**
> Q: Why will they resist our help? A: They will resist our help Due to excessive pride. **- Causal adjunct prepositional phrase as answer fragment**

No! It is not an example of short answer ☹:

> Q: Who walked the dog? A: Tom walked the dog.

4. Error:

Definition: error tag should only be annotated when there is a disagreement between the grounded knowledge graph and the utterances.

Example:

Yes! It is an example of short answer ☺:

> Q: How is the weather today? (However, the grounded relation is not talking about weather.)

No! It is not an example of short answer ☹:

> Q: When are Bill Gates bron? (Typo and grammatical error should not be annotated with error tags in our schema).

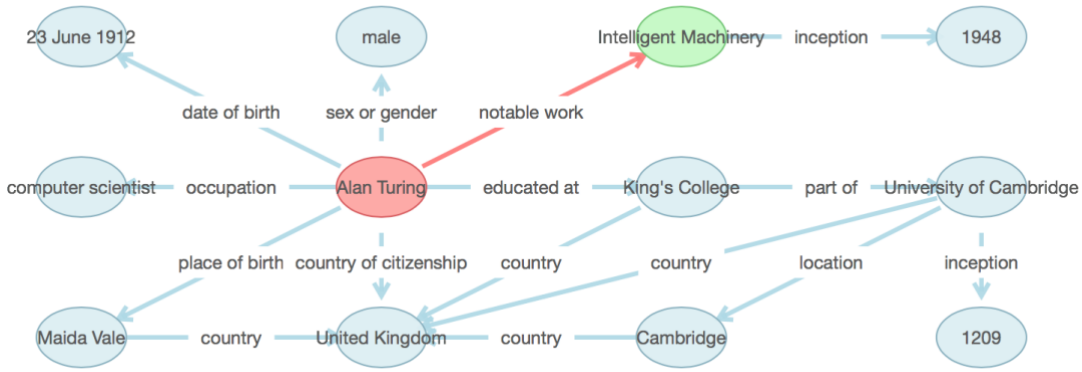
Labelling:

1. Additional information outside graph:

Definition: If parts of an utterance introduce additional information into the conversation which is not mentioned in the grounded knowledge graph, we call this part additional information outside the knowledge graph.

Example:

Given the knowledge graph and highlight:



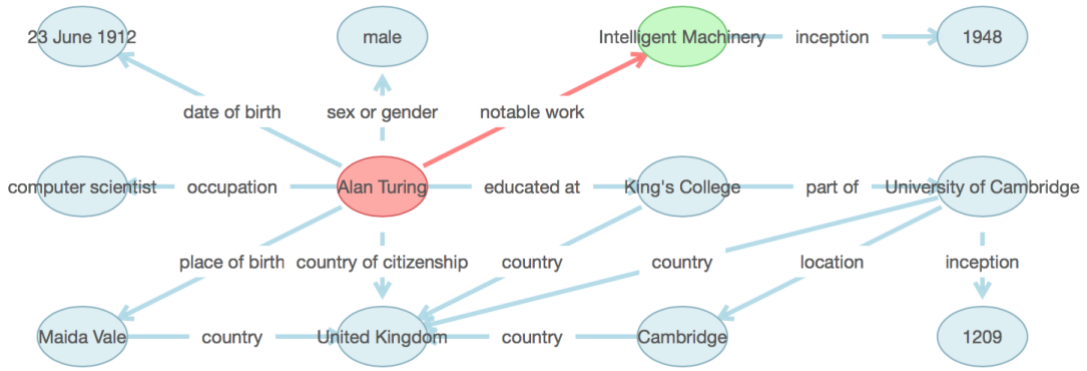
Yes! It is an example of Additional information outside graph ☺:

> Q: Can you tell me one of the notable work of Mr. Turning?
>
> A: Intelligent Machinery! Don't you head of the famous turning test?
>
> The underlining sentence is considered as addition information from outside of the knowledge graph, because there is no evidence that intelligent machinery is related to turning test.

2. Additional information within graph

Definition: If parts of an utterance introduce additional information into the conversation which is mentioned in the graph and such information is not highlighted, we call them additional information within the knowledge graph.

Given the knowledge graph and highlight:



Yes! It is an example of Additional information within graph ☺:

> Q: Can you tell me one of the notable work of Mr. Turning?
> A: Intelligent Machinery! Turing got this amazon idea in 1948.
> The underlining sentence is considered as addition information within of the knowledge graph, because the highlight relation doesn't contain year 1948. However, from the knowledge graph, we could know that Intelligent Machinery is incepted in 1948.

3. Conversation Control

Definition: Conversation Control label should be tagged to any pieces of utterance which functionally act as a connector to make the whole conversation more natural and coherent.

Yes! It is an example of Conversation Control ☺:

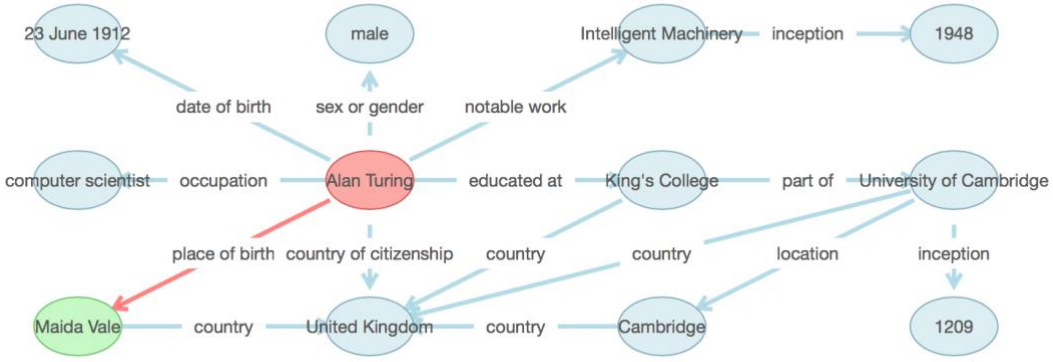
> Q: When was Alan Turing born?
> A: He was born in 1912.
> Q: And where?
> A: A town called Maida Vale.
> Q: Sorry, it is bit random. When was Alan's notable work, Intelligent Machinery incepted?
> A: 1948.

4. Totally Irrelevant

Definition: totally irrelevant tag should be used when pieces of utterance is totally irrelevant to the question answering conversation. By deleting such pieces of information, we will not influence the coherence of the conversation or the meaning of the utterance. In addition, this tag should only be considered

if the part of utterance is not labelled as other labels. In other words, labels, such as Conversation Control, will have higher precedence.

Example:

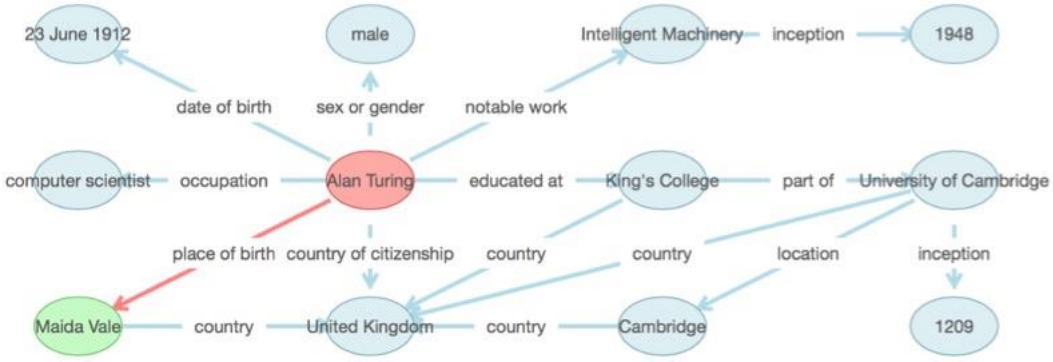


Yes! It is an example of Totally Irrelevant☺:

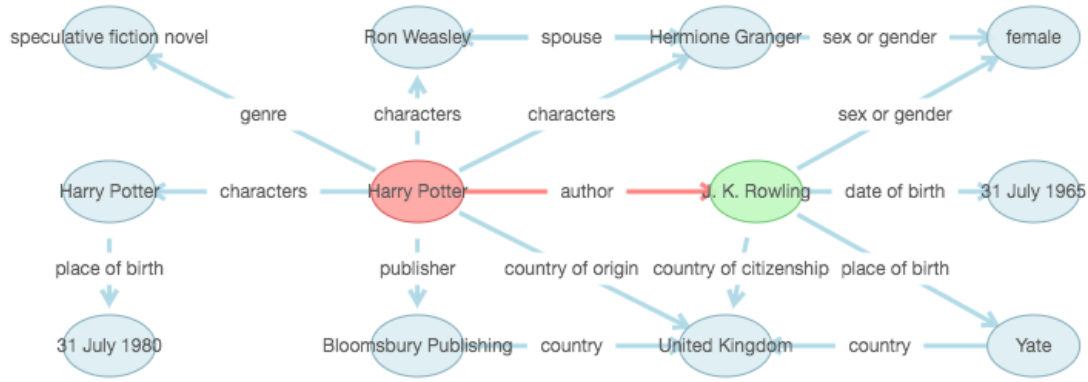
> 13,245,000. do you mind if we just do the next task and that's us finished or do you want to do another round?
>
> The underlining sentence is considered as totally irrelevant information since it is not directly related to our conversations.

Appendix A: All Grounded Graphs for this Tasks.

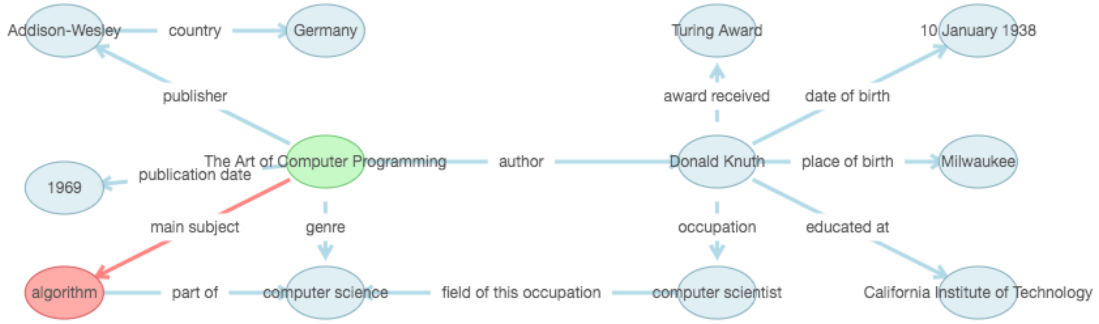
1.



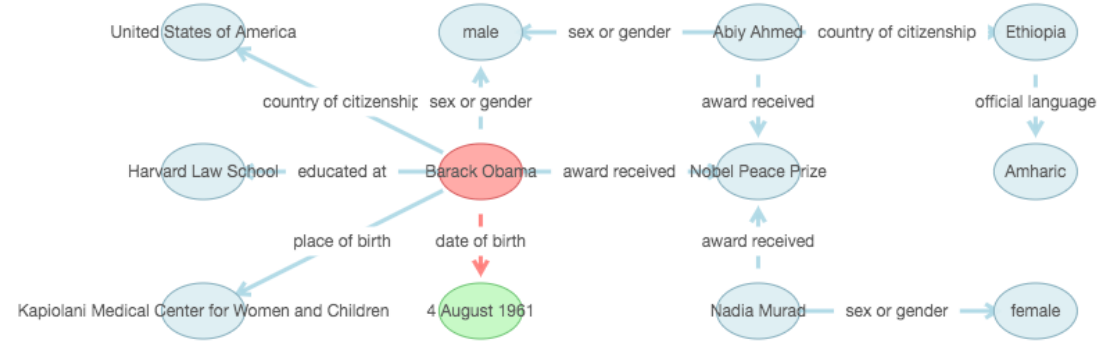
2.


speculative fiction novel
Ron Weasley
spouse
Hermione Granger
sex or gender
female
genre
characters
characters
sex or gender
Harry Potter
characters
Harry Potter
author
J. K. Rowling
date of birth
31 July 1965
place of birth
publisher
country of origin
country of citizenship
place of birth
31 July 1980
Bloomsbury Publishing
country
United Kingdom
country
Yate

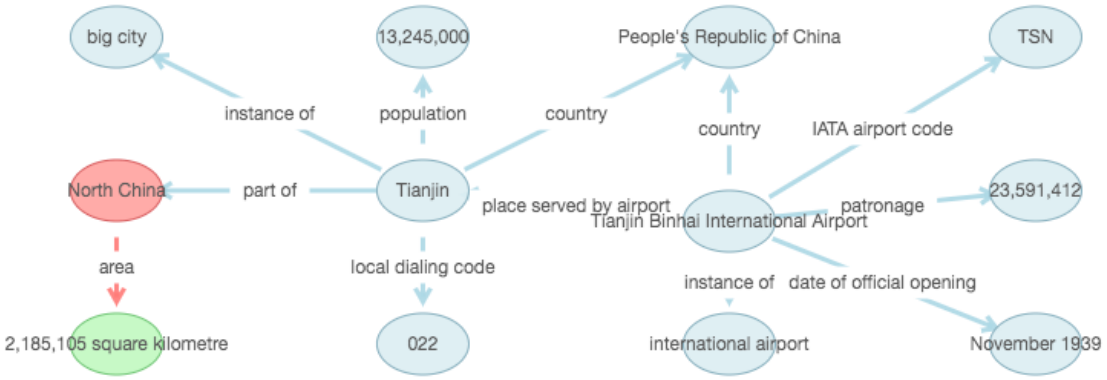
3.


Addison-Wesley
country
Germany
Turing Award
10 January 1938
publisher
award received
date of birth
1969
publication date
The Art of Computer Programming
author
Donald Knuth
place of birth
Milwaukee
main subject
genre
occupation
educated at
algorithm
part of
computer science
field of this occupation
computer scientist
California Institute of Technology

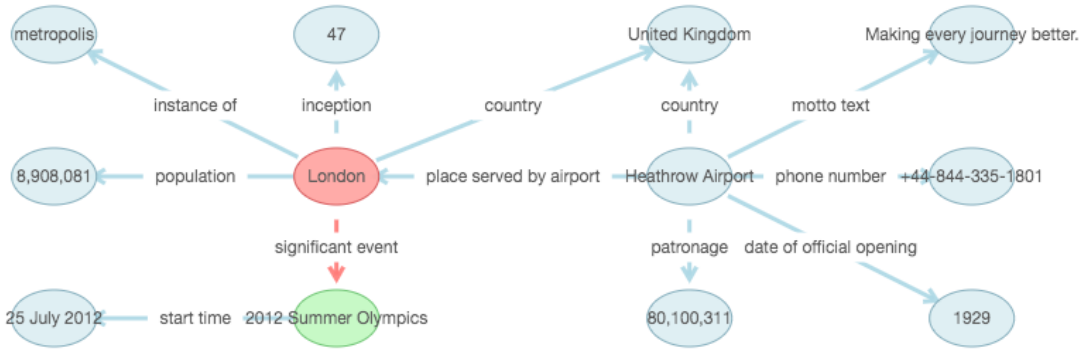
4.


United States of America
male
sex or gender
Abiy Ahmed
country of citizenship
Ethiopia
country of citizenship
sex or gender
award received
official language
Harvard Law School
educated at
Barack Obama
award received
Nobel Peace Prize
Amharic
place of birth
date of birth
award received
Kapiolani Medical Center for Women and Children
4 August 1961
Nadia Murad
sex or gender
female

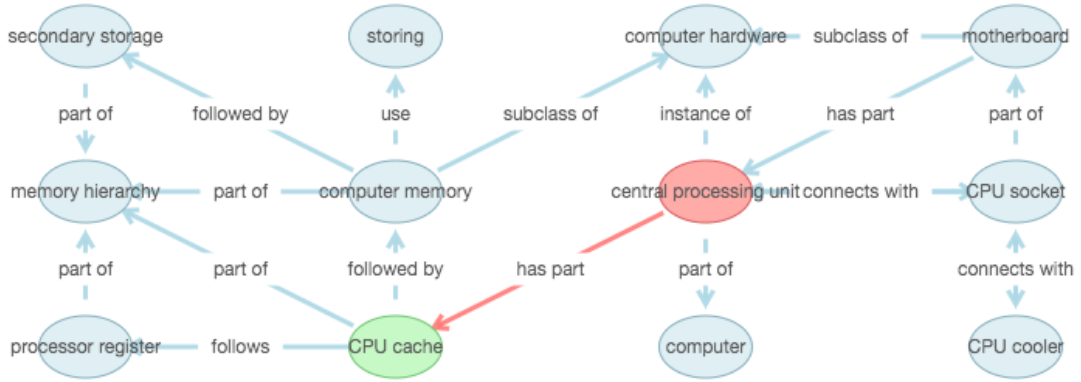
5.


big city
13,245,000
People's Republic of China
TSN
instance of
population
country
country
IATA airport code
North China
part of
Tianjin
place served by airport
Tianjin Binhai International Airport
patronage
23,591,412
area
local dialing code
instance of
date of official opening
2,185,105 square kilometre
022
international airport
November 1939

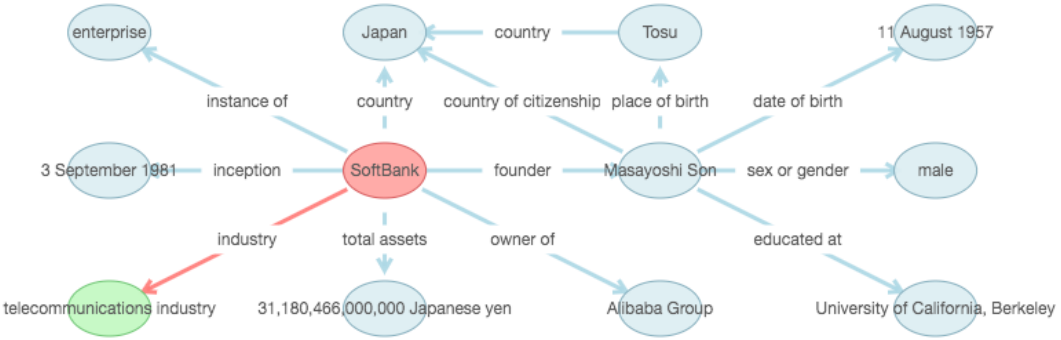
6.


metropolis
47
United Kingdom
Making every journey better.
instance of
inception
country
country
motto text
8,908,081
population
London
place served by airport
Heathrow Airport
phone number
+44-844-335-1801
significant event
patronage
date of official opening
25 July 2012
start time
2012 Summer Olympics
80,100,311
1929

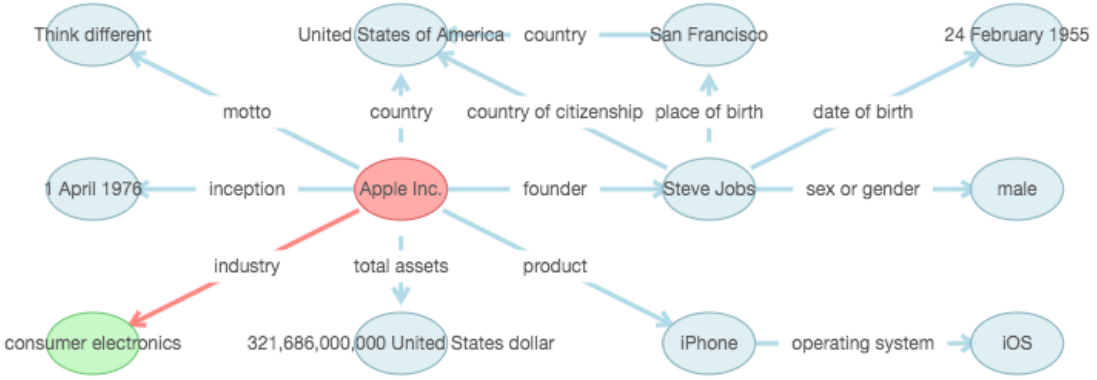
7.


secondary storage
storing
computer hardware
subclass of
motherboard
part of
followed by
use
subclass of
instance of
has part
part of
memory hierarchy
part of
computer memory
central processing unit
connects with
CPU socket
part of
part of
followed by
has part
part of
connects with
processor register
follows
CPU cache
computer
CPU cooler

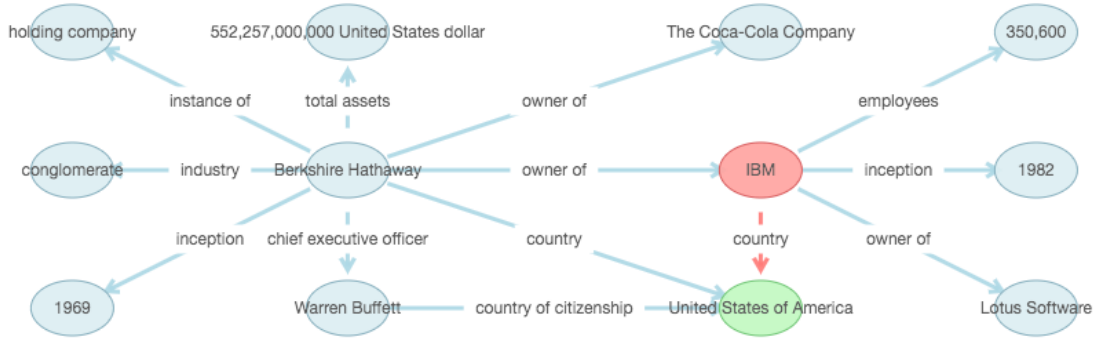
8.


enterprise
Japan
country
Tosu
11 August 1957
instance of
country
country of citizenship
place of birth
date of birth
3 September 1981
inception
SoftBank
founder
Masayoshi Son
sex or gender
male
industry
total assets
owner of
educated at
telecommunications industry
31,180,466,000,000 Japanese yen
Alibaba Group
University of California, Berkeley

9.

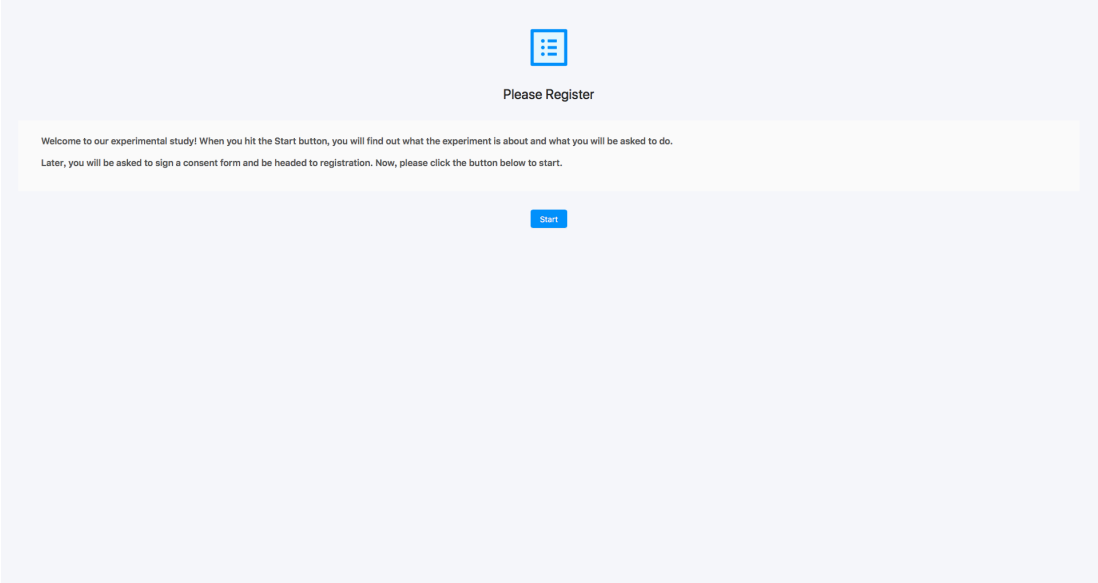

Think different
United States of America
country
San Francisco
24 February 1955
motto
country
country of citizenship
place of birth
date of birth
1 April 1976
inception
Apple Inc.
founder
Steve Jobs
sex or gender
male
industry
total assets
product
consumer electronics
321,686,000,000 United States dollar
iPhone
operating system
iOS

10.


holding company
552,257,000,000 United States dollar
The Coca-Cola Company
350,600
instance of
total assets
owner of
employees
conglomerate
industry
Berkshire Hathaway
owner of
IBM
inception
1982
inception
chief executive officer
country
country
owner of
1969
Warren Buffett
country of citizenship
United States of America
Lotus Software

Appendix B

Screenshots for Web-interface



Contact

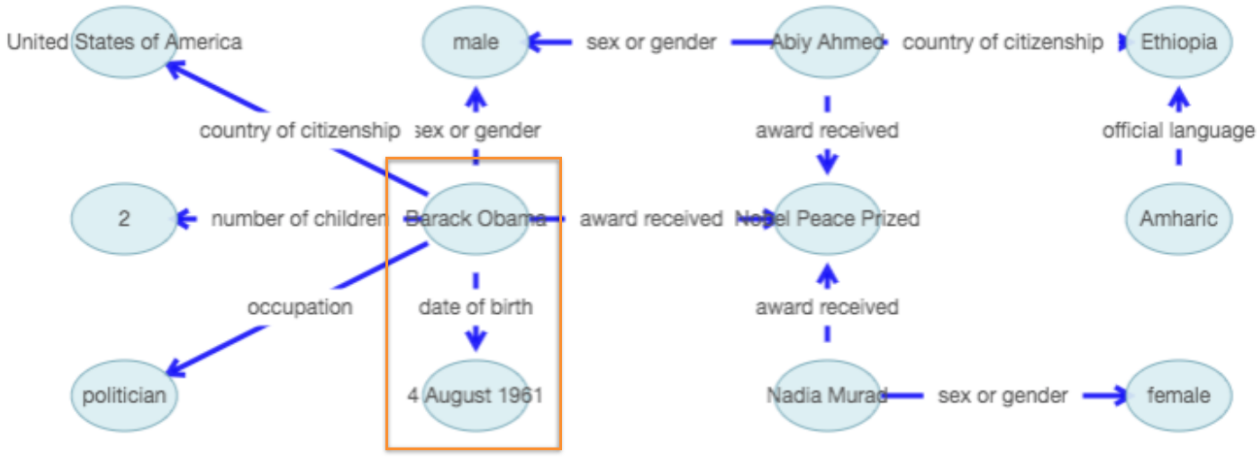
Songbo Hu
University of Edinburgh
s1647079@sms.ed.ac.uk

Introduction

The objective of this study is to collect conversations about the relations that are represented in a knowledge graph.
You have to participate in a natural conversation with your colleague in which you talk about the relations between entities as depicted in the knowledge graph. The knowledge graph will be displayed on your screen.

What is a Knowledge Graph?

A knowledge graph contains nodes that standing for individuals or things and arcs standing for the relations. The arrows on the arcs indicate the direction of the relation. Here is an example of a **knowledge graph**:

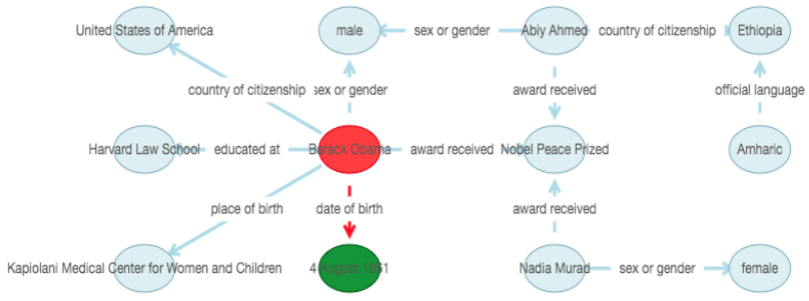


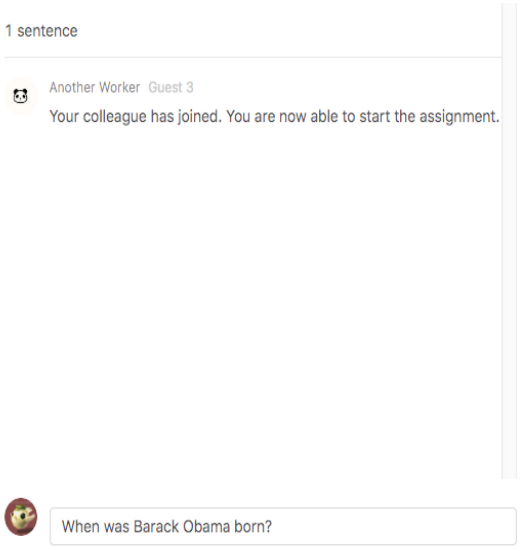
For instance, the relation in the **yellow box** above means Barack Obama was born in August 4, 1961.

Your task:

You and your colleague will each play a different role in the conversation. One of you is Questioner and the other is Answerer . You will talk in turns. You have freedom to choose how to phrase your sentences and keep the conversation natural.

Questioner is the initiator of the conversation. In order to start a conversation, Questioner should ask a **question** based on the **relation (arc)** shown in **RED** on the knowledge graph. The **node** marked in **GREEN** should be the **answer** to your question. For example:





In this case, Questioner should ask a question about the fact of date of birth of Obama which the answer to this question should be August 4, 1961.
For instance, Questioner might say any of the following, among others:

When was Barack Obama born?
or
What is the data of birth of Barack Obama?
or
When is Obama's birthday?
or
.....

All these questions are valid, but it is not an exhaustive list of what is valid. However,

When was the last president of the US born?

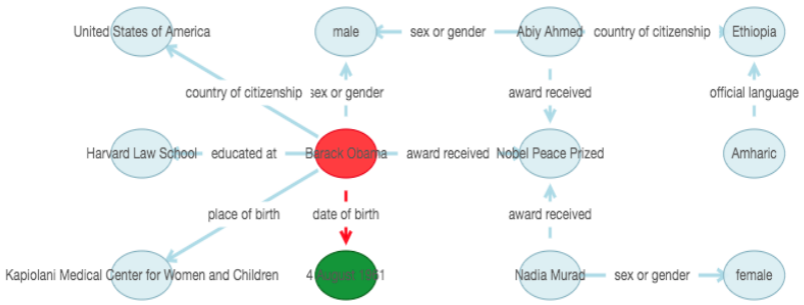
is not a valid question, because only from the highlighted relation, we wouldn't know that Barack Obama was the president of the US. We should not bring additional general knowledge to this conversation.

For example, Barack Obama was the president of the US would be considered as additional general knowledge, whereas, date of birth is the synonym of birthday is not additional general knowledge. Rephrasing is valid.

Then Questioner could send the question to Answerer and wait for the response.

Now it is the turn for Answerer 🐼 .

Answerer will receive the question from Questioner together with the same graph and the same colouring. In this case, it will be:

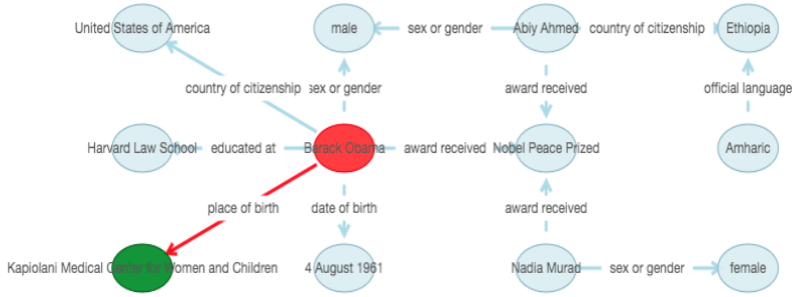


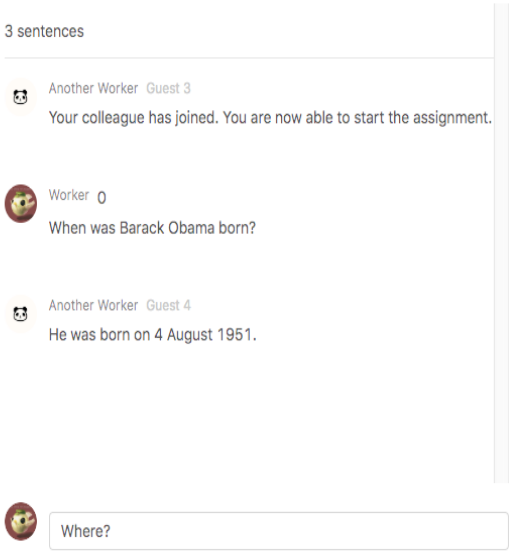
Answerer **may** reply:

He was born on 4 August 1951.
or
Obama was born on Aug 4th 1951.
or
4/Aug/1951.
or
.....

All these responses are valid. Again, it is not an exhaustive list of what is valid. Answerer now could send the answer back to Questioner and wait for the next question.

After Questioner receives the response from Answerer, the graph will be updated to highlight a new relation, together with the green node for the answer. Questioner should then ask another question about that relation, in a way that is natural given the conversation so far. For instance:

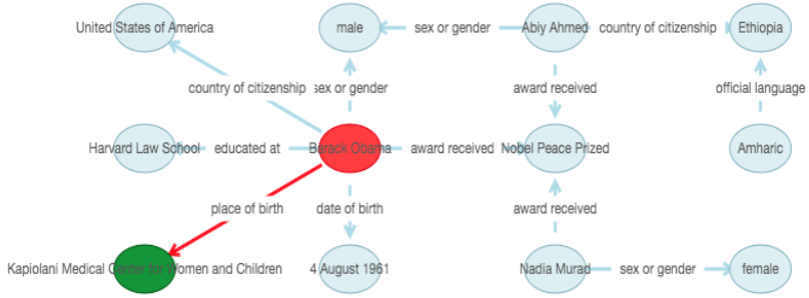


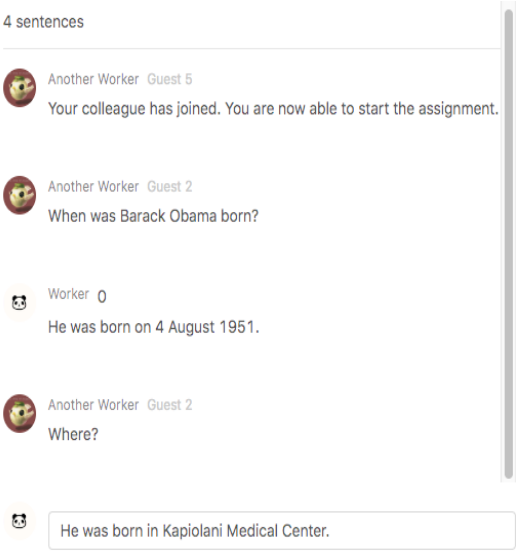


Where?
or
Where was Obama born?
or
Where was he born?
or
.....

In fact, in this context, any of the above can be a natural way to ask the question. Then, send the question to Answerer.

Answerer would again reply the question based on the updated graph and the context. Then send the response to Questioner.

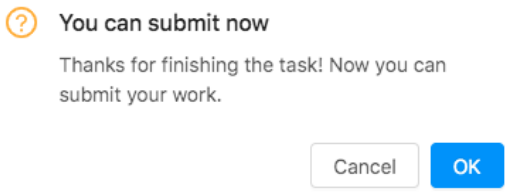




Again, there are many natural responses in this context. For example,

He was born in Kapiolani Medical Center.
or
In Kapiolani Medical Center.
or
.....

Keep going until the system prompt that you could submit their work. Normally it will takes 6 turns. You should click submit now after seeing the popup below.

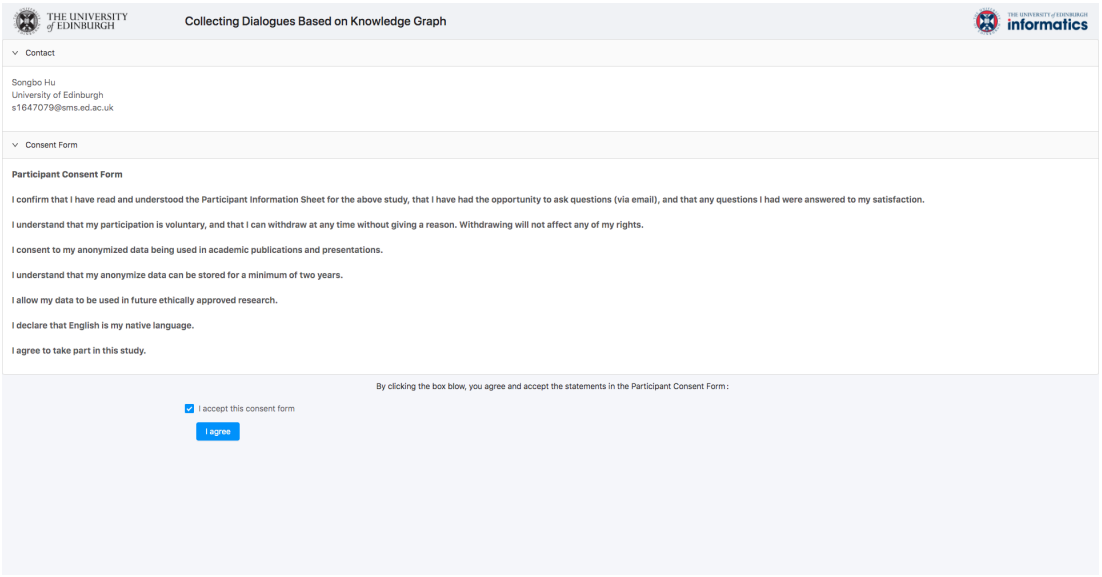


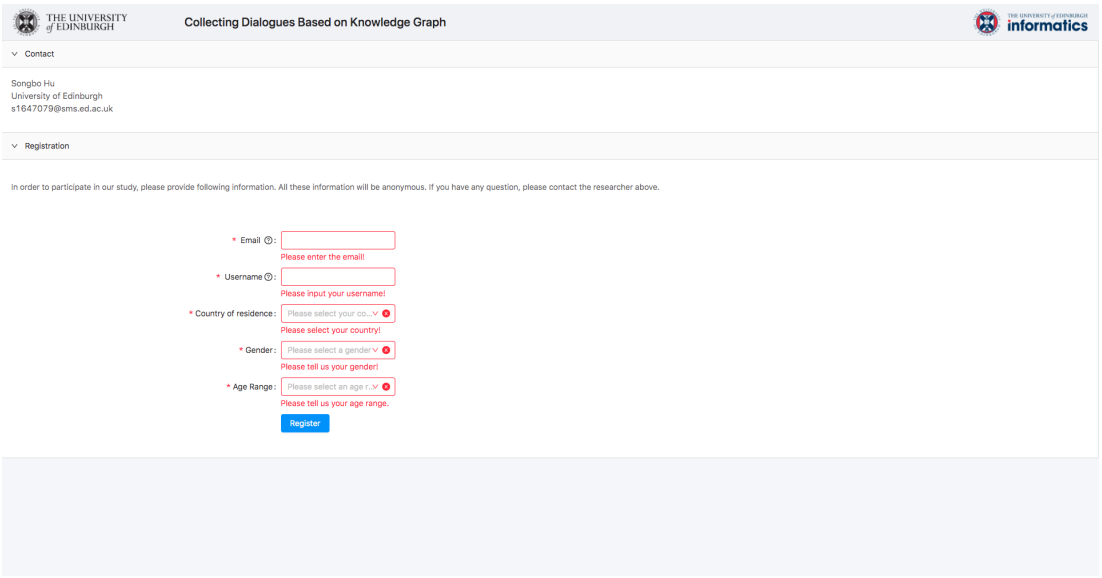
The following kinds of conversations are good because they are likely to happen in our daily life.

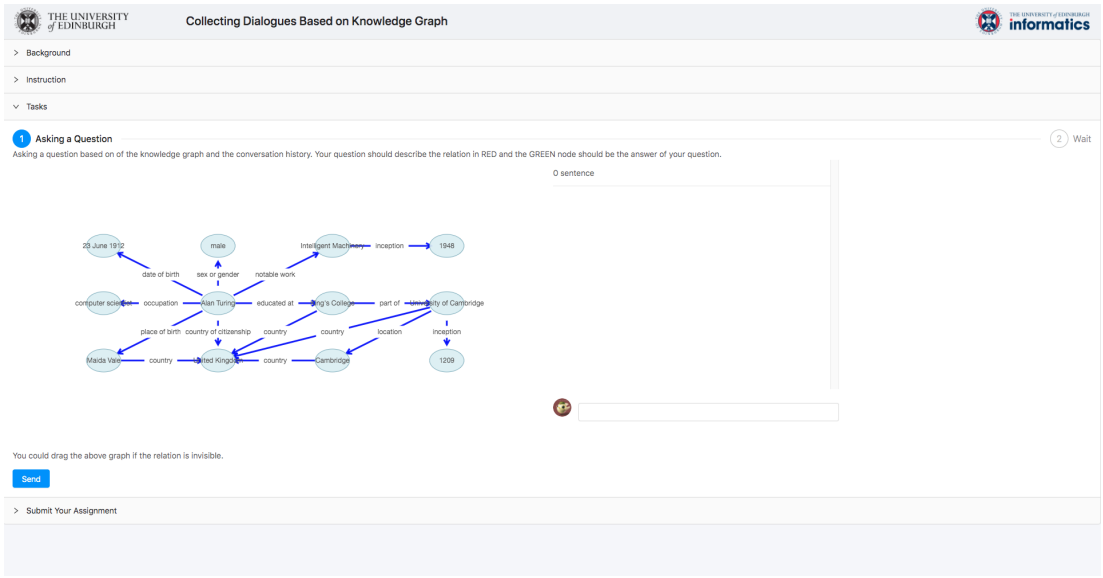
A: When was Barack Obama born?
B: 4 August 1951.
A: Where?
B: He was born in Kapiolani Medical Center.
A: Has he received any award?
B: Don't you remember this? He won the Nobel Peace Prized.
A: Wow! Who else has received this prize?
B: Nadia Murad.
A: Btw, I forgot to ask. Where is Obama from?
B: He's from the US.
.....

Click Next to continue.

Next


THE UNIVERSITY of EDINBURGH
Collecting Dialogues Based on Knowledge Graph
informatics
Contact
Songbo Hu
University of Edinburgh
s1647079@sms.ed.ac.uk
Consent Form
Participant Consent Form
I confirm that I have read and understood the Participant Information Sheet for the above study, that I have had the opportunity to ask questions (via email), and that any questions I had were answered to my satisfaction.
I understand that my participation is voluntary, and that I can withdraw at any time without giving a reason. Withdrawing will not affect any of my rights.
I consent to my anonymized data being used in academic publications and presentations.
I understand that my anonymize data can be stored for a minimum of two years.
I allow my data to be used in future ethically approved research.
I declare that English is my native language.
I agree to take part in this study.
By clicking the box blow, you agree and accept the statements in the Participant Consent Form:
I accept this consent form
I agree


THE UNIVERSITY of EDINBURGH
Collecting Dialogues Based on Knowledge Graph
informatics
Contact
Songbo Hu
University of Edinburgh
s1647079@sms.ed.ac.uk
Registration
In order to participate in our study, please provide following information. All these information will be anonymous. If you have any question, please contact the researcher above.
* Email:
Please enter the email!
* Username:
Please input your username!
* Country of residence:
Please select your co...
Please select your country!
* Gender:
Please select a gender
Please tell us your gender!
* Age Range:
Please select an age r...
Please tell us your age range.
Register


THE UNIVERSITY of EDINBURGH
Collecting Dialogues Based on Knowledge Graph
THE UNIVERSITY of EDINBURGH informatics
Background
Instruction
Tasks
1 Asking a Question
2 Wait
Asking a question based on of the knowledge graph and the conversation history. Your question should describe the relation in RED and the GREEN node should be the answer of your question.
0 sentence
23 June 1912
male
Intelligent Machinery
inception
1948
date of birth
sex or gender
notable work
computer scientist
occupation
Alan Turing
educated at
King's College
part of
University of Cambridge
place of birth
country of citizenship
country
country
location
inception
Maida Vale
country
United Kingdom
country
Cambridge
1209
You could drag the above graph if the relation is invisible.
Send
Submit Your Assignment

Appendix C

Leaflet for Advertising this Experiment

THE UNIVERSITY *of* EDINBURGH

informatics

PARTICIPANTS NEEDED

FOR RESEARCH PROJECT:

Dialogues Collecting Experiment

We are looking for **native English speakers** to participate in our experimental study. It takes approximately 45 minutes and you will be paid a **£5** Amazon voucher (from amazon.co.uk).

The objective of this study is to collect conversations about the relations that are represented in a knowledge graph. You will access a web-survey and participate in a natural conversation with one other participant.

If you are interested, please let me know by contacting:

s1647079@sms.ed.ac.uk

We will get in touch to arrange a time for your participation that is convenient for you. You may find additional details about this study via:

https://homepages.inf.ed.ac.uk/s1647079/index.html

Dialogues Collecting Experiment
s1647079@sms.ed.ac.uk
https://homepages.inf.ed.ac.uk/s1647079/index.html

Dialogues Collecting Experiment
s1647079@sms.ed.ac.uk
https://homepages.inf.ed.ac.uk/s1647079/index.html

Dialogues Collecting Experiment
s1647079@sms.ed.ac.uk
https://homepages.inf.ed.ac.uk/s1647079/index.html

Dialogues Collecting Experiment
s1647079@sms.ed.ac.uk
https://homepages.inf.ed.ac.uk/s1647079/index.html

Dialogues Collecting Experiment
s1647079@sms.ed.ac.uk
https://homepages.inf.ed.ac.uk/s1647079/index.html

Dialogues Collecting Experiment
s1647079@sms.ed.ac.uk
https://homepages.inf.ed.ac.uk/s1647079/index.html

Dialogues Collecting Experiment
s1647079@sms.ed.ac.uk
https://homepages.inf.ed.ac.uk/s1647079/index.html

Dialogues Collecting Experiment
s1647079@sms.ed.ac.uk
https://homepages.inf.ed.ac.uk/s1647079/index.html

Dialogues Collecting Experiment
s1647079@sms.ed.ac.uk
https://homepages.inf.ed.ac.uk/s1647079/index.html